NEWTON ITERATION, CONDITIONING AND ZERO COUNTING

GREGORIO MALAJOVICH

1. Introduction

Mathematicians' obsession with counting led to many interesting and far-fetched problems. These lectures are structured around a seemingly innocent counting problem:

Problem 1.1 (Real root counting)**.** Given a system $\mathbf{f} = (f_1, \ldots, f_n)$ of real polynomial equations in n variables, count the number of real solutions.

You can also find here a crash-course in Newton iteration. We will state and analyze a Newton iteration based 'inclusion-exclusion' algorithm to count (and find) roots of real polynomials.

That algorithm was investigated in a sequence of three papers by Felipe Cucker, Teresa Krick, Mario Wschebor and myself (2008, 2009, 2012). Good numerical properties are proved in the first paper. For instance, the algorithm is tolerant to controlled rounding error. Instead of covering such technicalities, I will present a simplified version and focus on the main ideas.

The interest of Problem 1.1 lies in the fact that it is **complete** for the complexity class $\#\mathbf{P}_{\mathbb{R}}$ over the **BSS** (Blum-Shub-Smale) computation model over $\mathbb{R}$. See Blum et al. (1998) for the BSS model of computation. The class $\#\mathbf{P}_{\mathbb{R}}$ was defined by Meer (2000) as the class of all functions $f : \mathbb{R}^\infty \to \{0,1\}^\infty \cup \{\infty\}$ such that there exists a BSS

Date: July 13, 2012.

Lecture notes for the Santaló summer school on Recent Advances in Real Complexity and Computation, held at the Palacio de la Magdalena, Santander, and sponsored by the Universidad Internacional Menéndez Pelayo and the Universidad de Cantábria.

G.M. is partially supported by CNPq and CAPES (Brazil) and by the Math-AmSud grant *complexity*.

machine M working in polynomial time and a polynomial q satisfying

$$f(\mathbf{y}) = \#\{\mathbf{z} \in \mathbb{R}^{q(\text{size}(\mathbf{y})} : M(\mathbf{y}, \mathbf{z}) \text{ is an accepting computation.}\}$$

We refer to Bürgisser and Cucker (2006) for the proof of completeness and to Cucker et al. (2008) for references on the subject of counting zeros.

Counting real polynomial roots in $\mathbb{R}^n$ can be reduced to counting polynomial roots in $\mathbb{S}^{n+1}$. Given a degree d polynomial $f(x_1, \dots, x_n)$, its homogenization is $f^{\text{homo}}(x_0, \dots, x_n) = x_0^d f(x_1/x_0, \dots, x_n/x_0)$.

Exercise 1.1 (Beware of infinity). Find an homogeneous polynomial $g = g(\mathbf{y}, u)$ of degree 2 in $n+2$ variables such that

$$\#\{\mathbf{x} \in \mathbb{R}^n : f_1(\mathbf{x}) = \cdots = f_n(\mathbf{x}) = 0\} + 1 =$$
$$= \frac{1}{2}\#\{(\mathbf{y}, u) \in \mathbb{S}^{n+1} : f_1^{\text{homo}}(\mathbf{y}) = \cdots = f_n^{\text{homo}}(\mathbf{y}) = g(\mathbf{y}, u) = 0\}.$$

Because of the exercise above, replacing n by $n-1$, Problem 1.1 reduces to:

Problem 1.2 (Real root counting on S^n)**.** Given a system $\mathbf{f} = (f_1, \dots, f_n)$ of real homogeneous polynomial equations in $n+1$ variables, count the number of solutions in S^n.

This course is organized as follows. We start by a review of **alpha-theory**. This theory originated with a couple of theorems proved by Steve Smale (1986) and improved subsequently by several authors. It allows to guarantee (quantitatively) from the available data that Newton iterations will converge quadratically to the solution of a system of equations.

Then I will speak about the inclusion-exclusion algorithm. It uses crucially several results of alpha-theory.

The complexity of the inclusion-exclusion algorithm depends upon a condition number. By endowing the input space with a probability distribution, one can speak of the expected value of the condition number and of the expected running time. The final section is a review of the complexity analysis performed in Cucker et al. (2009) and Cucker et al. (2012).

A warning: these lectures are informal. The model of computation is **cloud computing**. This means that we will allow for exponentially many parallel processors (essentially, BSS machines) at no additional cost. Moreover, we will be informal in the sense that we will assume that square roots and operator norms can be computed exactly in finite time. While this does not happen in the BSS model, those can be

approximated and all our algorithms can be rewritten as rigorous BSS algorithms at the cost of a harder complexity analysis (Cucker et al., 2008).

Exercise 1.2. What would happen if you could design a true polynomial time algorithm to solve Problem 1.2?

Acknowledgments. I would like to thank Teresa Krick, Felipe Cucker and Mike Shub for pointing out some mistakes in a previous version.

Contents

Part 1. Newton Iteration and Alpha theory

2. Outline

Let $\mathbf{f}$ be a mapping between Banach spaces. **Newton Iteration** is defined by

$$N(\mathbf{f}, \mathbf{x}) = \mathbf{x} - D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})$$

wherever $D\mathbf{f}(\mathbf{x})$ exists and is bounded. Its only possible fixed points are those satisfying $\mathbf{f}(\mathbf{x}) = 0$. When $\mathbf{f}(\mathbf{x}) = 0$ and $D\mathbf{f}(\mathbf{x})$ is invertible, we say that $\mathbf{x}$ is a **nondegenerate zero** of $\mathbf{f}$.

It is well-known that Newton iteration is quadratically convergent in a neighborhood of a nondegenerate zero ζ. Indeed, $N(\mathbf{f}, \mathbf{x}) - \zeta = D^2\mathbf{f}(\zeta)(\mathbf{x} - \zeta)^2 + \cdots$.

There are two main approaches to quantify how fast is quadratic convergence. One of them, pioneered by Kantorovich (1949) assumes that the mapping $\mathbf{f}$ has a bounded second derivative, and that this bound is known.

The other approach, developed by Smale (1985, 1986) and described here, assumes that the mapping $\mathbf{f}$ is analytic. Then we will be able to estimate a neighborhood of quadratic convergence around a given zero (Theorem 4.2) or to certify an 'approximate root' (Theorem 5.3) from data that depends only on the value and derivatives of $\mathbf{f}$ at one point.

A more general exposition on this subject may be found in Dedieu (1997b), covering also overdetermined and undetermined polynomial systems.

3. The gamma invariant

Through this chapter, $\mathbb{E}$ and $\mathbb{F}$ are Banach spaces, $\mathcal{D} \subseteq \mathbb{E}$ is open and $\mathbf{f} : \mathbb{E} \to \mathbb{F}$ is analytic.

This means that if $\mathbf{x}_0 \in \mathbb{E}$ is in the domain of $\mathbb{E}$, then there is $\rho > 0$ with the property that the series

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{D}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \mathbf{D}^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0, \mathbf{x} - \mathbf{x}_0) + \cdots \tag{1}$$

converges uniformly for $\|\mathbf{x} - \mathbf{x}_0\| < \rho$, and its limit is equal to $\mathbf{f}(\mathbf{x})$ (For more details about analytic functions between Banach spaces, see Nachbin (1964, 1969)).

In order to abbreviate notations, we will write (1) as

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{D}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \sum_{k \geq 2} \frac{1}{k!} \mathbf{D}^k f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)^k$$

where the exponent k means that $\mathbf{x} - \mathbf{x}_0$ appears k times as an argument to the preceding multi-linear operator.

The maximum of such ρ will be called the **radius of convergence**. (It is ∞ when the series (1) is globally convergent). This terminology comes from univariate complex analysis. When $\mathbf{E} = \mathbb{C}$, the series will converge for all $\mathbf{x} \in B(\mathbf{x}_0, \rho)$ and diverge for all $\mathbf{x} \notin \overline{B(\mathbf{x}_0, \rho)}$. This is no more true in several complex variables, or Banach spaces (Exercise 4.1).

The norm of a k-linear operator in Banach Spaces (such as the k-th derivative) is the **operator norm**, for instance

$$\|D^k\mathbf{f}(\mathbf{x}_0)\|_{\mathbb{E}\to\mathbb{F}} = \sup_{\|\mathbf{u}_1\|_{\mathbb{E}}=\cdots=\|\mathbf{u}_k\|_{\mathbb{E}}=1} \|D^k\mathbf{f}(\mathbf{x}_0)(\mathbf{u}_1,\ldots,\mathbf{u}_k)\|_{\mathbb{F}}.$$

As long as there is no ambiguity, we drop the subscripts of the norm.

Definition 3.1 (Smale's γ invariant)**.** Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic mapping between Banach spaces, and $\mathbf{x}_0 \in \mathcal{D}$. When $D\mathbf{f}(\mathbf{x}_0)$ is invertible, define

$$\gamma(\mathbf{f},\mathbf{x}_0) = \sup_{k\geq 2}\left(\frac{\|D\mathbf{f}(\mathbf{x}_0)^{-1}D^k\mathbf{f}(\mathbf{x}_0)\|}{k!}\right)^{\frac{1}{k-1}}.$$

Otherwise, set $\gamma(\mathbf{f},\mathbf{x}_0) = \infty$.

In the one variable setting, this can be compared to the radius of convergence ρ of $\mathbf{f}'(\mathbf{x})/\mathbf{f}'(\mathbf{x}_0)$, that satisfies

$$\rho^{-1} = \limsup_{k\geq 2}\left(\frac{\|\mathbf{f}'(\mathbf{x}_0)^{-1}\mathbf{f}^{(k)}(\mathbf{x}_0)\|}{k!}\right)^{\frac{1}{k-1}}.$$

More generally,

Proposition 3.2. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be a C^∞ map between Banach spaces, and $\mathbf{x}_0 \in \mathcal{D}$. Then f is analytic in x_0 if and only if, $\gamma(f,x_0)$ is finite. The series*

$$\mathbf{f}(\mathbf{x}_0) + \mathbf{D}f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \sum_{k\geq 2}\frac{1}{k!}\mathbf{D}^k f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^k \tag{2}$$

is uniformly convergent for $\mathbf{x} \in B(\mathbf{x}_0,\rho)$ for any $\rho < 1/\gamma(\mathbf{f},\mathbf{x}_0))$.

Proof of the **if** *in Prop.3.2.* The series

$$\mathbf{D}f(\mathbf{x}_0)^{-1}\mathbf{f}(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0) + \sum_{k\geq 2}\frac{1}{k!}\mathbf{D}f(\mathbf{x}_0)^{-1}\mathbf{D}^k f(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^k$$

is uniformly convergent in $B(\mathbf{x}_0,\rho)$ where

$$\begin{aligned}
\rho^{-1} &< \limsup_{k\geq 2}\left(\frac{\|D\mathbf{f}(\mathbf{x}_0)^{-1}D^k\mathbf{f}(\mathbf{x}_0)\|}{k!}\right)^{\frac{1}{k}} \\
&\leq \limsup_{k\geq 2}\gamma(\mathbf{f},\mathbf{x}_0)^{\frac{k-1}{k}} \\
&= \lim_{k\to\infty}\gamma(\mathbf{f},\mathbf{x}_0)^{\frac{k-1}{k}} \\
&= \gamma(\mathbf{f},\mathbf{x}_0)
\end{aligned}$$

□

Before proving the **only if** part of Proposition 3.2, we need to relate the norm of a multi-linear map to the norm of the corresponding polynomial.

Lemma 3.3. *Let $k \geq 2$. Let $\mathbf{T} : \mathbb{E}^k \to \mathbb{F}$ be k-linear and symmetric. Let $\mathbf{S} : \mathbb{E} \to \mathbb{F}$, $\mathbf{S}(\mathbf{x}) = T(\mathbf{x}, \mathbf{x}, \ldots, \mathbf{x})$ be the corresponding polynomial. Then,*

$$\|\mathbf{T}\| \leq e^{k-1} \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{S}(\mathbf{x})\|$$

Proof. The polarization formula for (real or complex) tensors is

$$\mathbf{T}(\mathbf{x}_1, \cdots, \mathbf{x}_k) = \frac{1}{2^k k!} \sum_{\substack{\epsilon_j = \pm 1 \\ j=1,\ldots,k}} \epsilon_1 \cdots \epsilon_k \mathbf{S}\left(\sum_{l=1}^{k} \epsilon_l \mathbf{x}_l \right)$$

It is easily derived by expanding the expression inside parentheses. There will be $2^k k!$ terms of the form

$$\epsilon_1 \cdots \epsilon_k T(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k)$$

or its permutations. All other terms miss at least one variable (say $\mathbf{x}_j$). They cancel by summing for $\epsilon_j = \pm 1$.

It follows that when $\|\mathbf{x}\| \leq 1$,

$$\begin{aligned} \mathbf{T}(\mathbf{x}_1, \cdots, \mathbf{x}_k) &\leq \frac{1}{k!} \max_{\substack{\epsilon_j = \pm 1 \\ j=1,\ldots,k}} \left\| \mathbf{S}\left(\sum_{l=1}^{k} \epsilon_l \mathbf{x}_l \right) \right\| \\ &\leq \frac{k^k}{k!} \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{S}(\mathbf{x})\| \end{aligned}$$

The Lemma follows from using Stirling's formula,

$$k! \geq \sqrt{2\pi k} k^k e^{-k} e^{1/(12k+1)}.$$

We obtain:

$$\|\mathbf{T}\| \leq \left(\frac{1}{\sqrt{2\pi k}} e^{-\frac{1}{12k+1}} \right) e^k \sup_{\|\mathbf{x}\| \leq 1} \|\mathbf{S}(\mathbf{x})\|.$$

Then we use the fact that $k \geq 2$, hence $\sqrt{2\pi k} \geq e$. □

Proof of Prop.3.2, **only if** *part.* Assume that the series (2) converges uniformly for $\|\mathbf{x} - \mathbf{x}_0\| < \rho$. Without loss of generality assume that $\mathbb{E} = \mathbb{F}$ and $D\mathbf{f}(\mathbf{x}_0) = I$.

We claim that

$$\limsup_{k \geq 2} \sup_{\|\mathbf{u}\|=1} \| \frac{1}{k!} D^k \mathbf{f}(\mathbf{x}_0) \mathbf{u}^k \|^{1/k} \leq \rho^{-1}.$$

Indeed, assume that there is $\delta > 0$ and infinitely many pairs $(k, \mathbf{u})$ with $\|\mathbf{u}_i\| = 1$ and

$$\|\frac{1}{k!}D^k\mathbf{f}(\mathbf{x}_0)\mathbf{u}^k\|^{1/k} > \rho^{-1}(1+\delta).$$

In that case,

$$\|\frac{1}{k!}D^k\mathbf{f}(\mathbf{x}_0)\left(\frac{\rho}{\sqrt{1+\delta}}\mathbf{u}\right)^k\| > \left(\sqrt{1+\delta}\right)^k$$

infinitely many times, and hence (2) does not converge uniformly on $B(\mathbf{x}_0, \rho)$.

Now, we can apply Lemma 3.3 to obtain:

$$\begin{aligned}\limsup_{k\geq 2}\|\frac{1}{k!}D^k\mathbf{f}(\mathbf{x}_0)\|^{1/(k-1)} &\leq e\limsup_{k\geq 2}\sup_{\|\mathbf{u}\|=1}\|\frac{1}{k!}D^k\mathbf{f}(\mathbf{x}_0)\mathbf{u}^k\|^{\frac{1}{k-1}}\\ &\leq e\lim_{k\to\infty}\rho^{-(1+1/(k-1))}\\ &= e\rho^{-1}\end{aligned}$$

and therefore $\|\frac{1}{k!}D^k f(x_0)\|^{1/(k-1)}$ is bounded. □

Exercise 3.1. Show the polarization formula for Hermitian product:

$$\langle\mathbf{u},\mathbf{v}\rangle = \frac{1}{4}\sum_{\epsilon^4=1}\epsilon\|\mathbf{u}+\epsilon\mathbf{v}\|^2$$

Explain why this is different from the one in Lemma 3.3.

Exercise 3.2. If one drops the uniform convergence hypothesis in the definition of analytic functions, what happens to Proposition 3.2?

4. The γ-Theorems

The following concept provides a good abstraction of quadratic convergence.

Definition 4.1 (Approximate zero of the first kind)**.** Let $\mathbf{f} : \mathcal{D} \subseteq \mathbf{E} \to \mathbf{F}$ be as above, with $\mathbf{f}(\zeta) = 0$. An **approximate zero of the first kind** associated to ζ is a point $\mathbf{x}_0 \in \mathcal{D}$, such that

(1) The sequence $(\mathbf{x})_i$ defined inductively by $\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i)$ is well-defined (each $\mathbf{x}_i$ belongs to the domain of $\mathbf{f}$ and $D\mathbf{f}(\mathbf{x}_i)$ is invertible and bounded).

(2)
$$\|\mathbf{x}_i - \zeta\| \leq 2^{-2^i+1}\|\mathbf{x}_0 - \zeta\|.$$

The existence of approximate zeros of the first kind is not obvious, and requires a theorem.

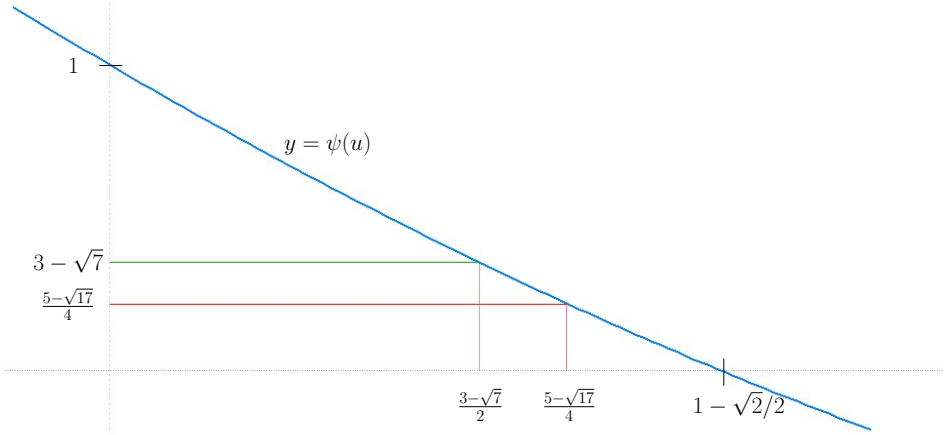

FIGURE 1. $y = \psi(u)$

Theorem 4.2 (Smale)**.** *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let ζ be a nondegenerate zero of $\mathbf{f}$. Assume that*

$$B = B\left(\zeta, \frac{3-\sqrt{7}}{2\gamma(\mathbf{f}, \zeta)}\right) \subseteq \mathcal{D}.$$

Every $\mathbf{x}_0 \in B$ is an approximate zero of the first kind associated to ζ. The constant $(3-\sqrt{7})/2$ is the smallest with that property.

Before going further, we remind the reader of the following fact.

Lemma 4.3. *Let $d \geq 1$ be integer, and let $|t| < 1$. Then,*

$$\frac{1}{(1-t)^d} = \sum_{k \geq 0} \binom{k+d-1}{d-1} t^k.$$

Proof. Differentiate $d-1$ times the two sides of the expression $1/(1-t) = 1 + t + t^2 + \cdots$, and then divide both sides by $d-1!$ □

Lemma 4.4. *The function $\psi(u) = 1 - 4u + 2u^2$ is decreasing and non-negative in $[0, 1-\sqrt{2}/2]$, and satisfies:*

$$\frac{u}{\psi(u)} < 1 \qquad \text{for } u \in [0, (5-\sqrt{17})/4) \tag{3}$$

$$\frac{u}{\psi(u)} \leq \frac{1}{2} \qquad \text{for } u \in [0, (3-\sqrt{7})/2]\ . \tag{4}$$

The proof of Lemma 4.4 is left to the reader (but see Figure 1).

Another useful result is:

Lemma 4.5. *Let A be a $n \times n$ matrix. Assume $\|A - I\|_2 < 1$. Then A has full rank and, for all y,*

$$\frac{\|y\|}{1 + \|A - I\|_2} \leq \|A^{-1}y\|_2 \leq \frac{\|y\|}{1 - \|A - I\|_2}.$$

Proof. By hypothesis, $\|Ax\| > 0$ for all $x \neq 0$ so that A has full rank. Let $y = Ax$. By triangular inequality,

$$\|Ax\| \geq \|x\| - \|(A - I)x\| \geq (1 - \|(A - I)\|_2)\|x\|.$$

Also by triangular inequality,

$$\|Ax\| \leq \|x\| + \|(A - I)x\| \leq (1 + \|(A - I)\|_2)\|x\|.$$

□

The following Lemma will be needed:

Lemma 4.6. *Assume that $u = \|\mathbf{x} - \mathbf{y}\|\gamma(\mathbf{f}, \mathbf{x}) < 1 - \sqrt{2}/2$. Then,*

$$\|D\mathbf{f}(\mathbf{y})^{-1}D\mathbf{f}(\mathbf{x})\| \leq \frac{(1-u)^2}{\psi(u)}.$$

Proof. Expanding $\mathbf{y} \mapsto D\mathbf{f}(\mathbf{x})^{-1}D\mathbf{f}(\mathbf{y})$ around $\mathbf{x}$, we obtain:

$$D\mathbf{f}(\mathbf{x})^{-1}D\mathbf{f}(\mathbf{y}) = I + \sum_{k \geq 2} \frac{1}{k-1!} D\mathbf{f}(\mathbf{x})^{-1}D^k\mathbf{f}(\mathbf{x})(\mathbf{y} - \mathbf{x})^{k-1}.$$

Rearranging terms and taking norms, Lemma 4.3 yields

$$\|D\mathbf{f}(\mathbf{x})^{-1}D\mathbf{f}(\mathbf{y}) - I\| \leq \frac{1}{(1 - \gamma\|\mathbf{y} - \mathbf{x}\|)^2} - 1.$$

By Lemma 4.5 we deduce that $D\mathbf{f}(\mathbf{x})^{-1}D\mathbf{f}(\mathbf{y})$ is invertible, and

$$\|D\mathbf{f}(\mathbf{y})^{-1}D\mathbf{f}(\mathbf{x})\| \leq \frac{1}{1 - \|D\mathbf{f}(\mathbf{x})^{-1}D\mathbf{f}(\mathbf{y}) - I\|} = \frac{(1-u)^2}{\psi(u)}. \tag{5}$$

□

Here is the method for proving Theorem 4.2 and similar ones: first we study the convergence of Newton iteration applied to a 'universal' function. In this case, set

$$h_\gamma(t) = t - \gamma t^2 - \gamma^2 t^3 - \cdots = t - \frac{\gamma t^2}{1 - \gamma t}.$$

(See figure 2).

The function h_γ has a zero at $t = 0$, and $\gamma(h_\gamma, 0) = \gamma$. Then, we compare the convergence of Newton iteration applied to an arbitrary function to the convergence when applied to the universal function.

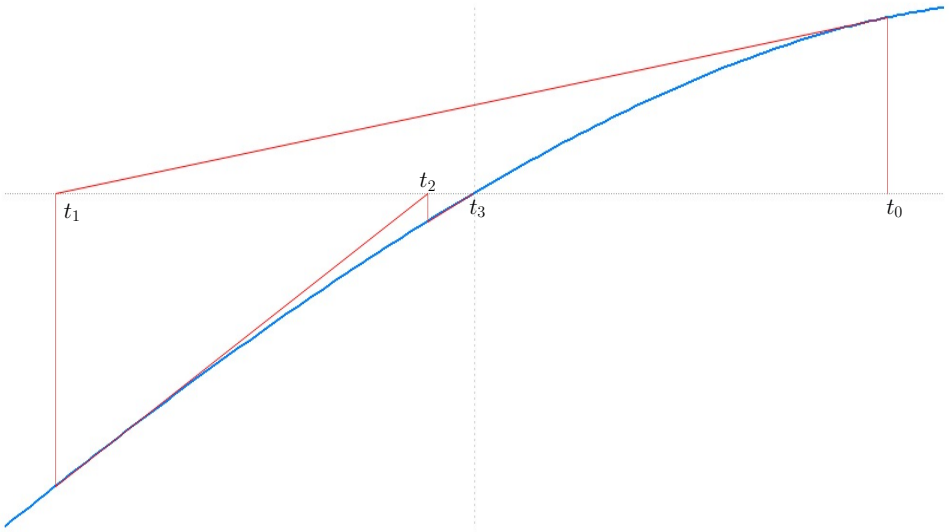

FIGURE 2. $y = h_\gamma(t)$

Lemma 4.7. *Assume that* $0 \leq u_0 = \gamma t_0 < \frac{5-\sqrt{17}}{4}$. *Then the sequences*

$$t_{i+1} = N(h_\gamma, t_i) \text{ and } u_{i+1} = \frac{u_i^2}{\psi(u_i)}$$

are well-defined for all i, $\lim_{i\to\infty} t_i = 0$, *and*

$$\frac{|t_i|}{|t_0|} = \frac{u_i}{u_0} \leq \left(\frac{u_0}{\psi(u_0)}\right)^{2^i-1}.$$

Moreover,

$$\frac{|t_i|}{|t_0|} \leq 2^{-2^i+1}$$

for all i *if and only if* $u_0 \leq \frac{3-\sqrt{7}}{2}$.

Proof. We just compute

$$\begin{aligned} h'_\gamma(t) &= \frac{\psi(\gamma t)}{(1-\gamma t)^2} \\ t h'_\gamma(t) - h_\gamma(t) &= -\frac{\gamma t^2}{(1-\gamma t)^2} \\ N(h_\gamma, t) &= -\frac{\gamma t^2}{\psi(\gamma t)}. \end{aligned}$$

When $u_0 < \frac{5-\sqrt{17}}{4}$, (3) implies that the sequence u_i is decreasing, and by induction

$$u_i = \gamma|t_i|.$$

Moreover,

$$\frac{u_{i+1}}{u_0} = \left(\frac{u_i}{u_0}\right)^2 \frac{u_0}{\psi(u_i)} \leq \left(\frac{u_i}{u_0}\right)^2 \frac{u_0}{\psi(u_0)} < \left(\frac{u_i}{u_0}\right)^2.$$

By induction,

$$\frac{u_i}{u_0} \leq \left(\frac{u_0}{\psi(u_0)}\right)^{2^i-1}.$$

This also implies that $\lim t_i = 0$.

When furthermore $u_0 \leq (3-\sqrt{7})/2$, $u_0/\psi(u_0) \leq 1/2$ by (4) hence $u_i/u_0 \leq 2^{-2^i+1}$. For the converse, if $u_0 > (3-\sqrt{7})/2$, then

$$\frac{|t_1|}{|t_0|} = \frac{u_0}{\psi(u_0)} > \frac{1}{2}.$$

□

Before proceeding to the proof of Theorem 4.2, a remark is in order.

Both Newton iteration and γ are invariant with respect to translation and to linear changes of coordinates: let $\mathbf{g}(\mathbf{x}) = A\mathbf{f}(\mathbf{x}-\zeta)$, where A is a continuous and invertible linear operator from $\mathbb{F}$ to $\mathbb{E}$. Then

$$N(\mathbf{g}, \mathbf{x}+\zeta) = N(\mathbf{f}, \mathbf{x}) + \zeta \text{ and } \gamma(\mathbf{g}, \mathbf{x}+\zeta) = \gamma(\mathbf{f}, \mathbf{x}).$$

Also, distances in $\mathbb{E}$ are invariant under translation.

Proof of Th.4.2. Assume without loss of generality that $\zeta = 0$ and $D\mathbf{f}(\zeta) = I$. Set $\gamma = \gamma(\mathbf{f}, \mathbf{x})$, $u_0 = \|\mathbf{x}_0\|\gamma$, and let h_γ and the sequence (u_i) be as in Lemma 4.7.

We will bound

$$\|N(\mathbf{f}, \mathbf{x})\| = \left\|\mathbf{x} - D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})\right\| \leq \|D\mathbf{f}(\mathbf{x})^{-1}\|\|\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{x}\|. \tag{6}$$

The Taylor expansions of $\mathbf{f}$ and $D\mathbf{f}$ around 0 are respectively:

$$\mathbf{f}(\mathbf{x}) = \mathbf{x} + \sum_{k\geq 2} \frac{1}{k!} D^k\mathbf{f}(0)\mathbf{x}^k$$

and

$$D\mathbf{f}(\mathbf{x}) = I + \sum_{k\geq 2} \frac{1}{k-1!} D^k\mathbf{f}(0)\mathbf{x}^{k-1}. \tag{7}$$

Combining the two equations, above, we obtain:

$$\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{x} = \sum_{k\geq 2} \frac{k-1}{k!} D^k\mathbf{f}(0)\mathbf{x}^k.$$

Using Lemma 4.3 with $d = 2$, the rightmost term in (6) is bounded above by

$$\|\mathbf{f}(\mathbf{x}) - D\mathbf{f}(\mathbf{x})\mathbf{x}\| \leq \sum_{k\geq 2}(k-1)\gamma^{k-1}\|\mathbf{x}\|^k = \frac{\gamma\|\mathbf{x}\|^2}{(1-\gamma\|\mathbf{x}\|)^2}. \tag{8}$$

Combining Lemma 4.6 and (8) in (6), we deduce that

$$\|N(\mathbf{f},\mathbf{x})\| \leq \frac{\gamma\|\mathbf{x}\|^2}{\psi(\gamma\|\mathbf{x}\|)}.$$

By induction, $u_i \leq \gamma\|\mathbf{x}_i\|$. When $u_0 \leq (3-\sqrt{7})/2$, we obtain as in Lemma 4.7 that

$$\frac{\|\mathbf{x}_i\|}{\|\mathbf{x}_0\|} \leq \frac{u_i}{u_0} \leq 2^{-2^i+1}.$$

We have seen in Lemma 4.7 that the bound above fails for $i = 1$ when $u_0 > (3-\sqrt{7})/2$. □

Notice that in the proof above,

$$\lim_{i\to\infty} \frac{u_0}{\psi(u_i)} = u_0.$$

Therefore, convergence is actually faster than predicted by the definition of approximate zero. We proved actually a sharper result:

Theorem 4.8. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let ζ be a nondegenerate zero of $\mathbf{f}$. Let $u_0 < (5-\sqrt{17})/4$.*

Assume that

$$B = B\left(\zeta, \frac{u_0}{\gamma(\mathbf{f},\zeta)}\right) \subseteq \mathcal{D}.$$

If $\mathbf{x}_0 \in B$, then the sequences

$$\mathbf{x}_{i+1} = N(\mathbf{f},\mathbf{x}_i) \text{ and } u_{i+1} = \frac{u_i^2}{\psi(u_i)}$$

are well-defined for all i, and

$$\frac{\|\mathbf{x}_i - \zeta\|}{\|\mathbf{x}_0 - \zeta\|} \leq \frac{u_i}{u_0} \leq \left(\frac{u_0}{\psi(u_0)}\right)^{-2^i+1}.$$

Table 1 and Figure 3 show how fast u_i/u_0 decreases in terms of u_0 and i.

To conclude this section, we need to address an important issue for numerical computations. Whenever dealing with digital computers, it is convenient to perform calculations in floating point format. This means that each real number is stored as a **mantissa** (an integer,

	1/32	1/16	1/10	1/8	$\frac{3-\sqrt{7}}{2}$
1	4.810	3.599	2.632	2.870	1.000
2	14.614	11.169	8.491	6.997	3.900
3	34.229	26.339	20.302	16.988	10.229
4	73.458	56.679	43.926	36.977	22.954
5	151.917	117.358	91.175	76.954	48.406

TABLE 1. Values of $-log_2(u_i/u_0)$ in function of u_0 and i.

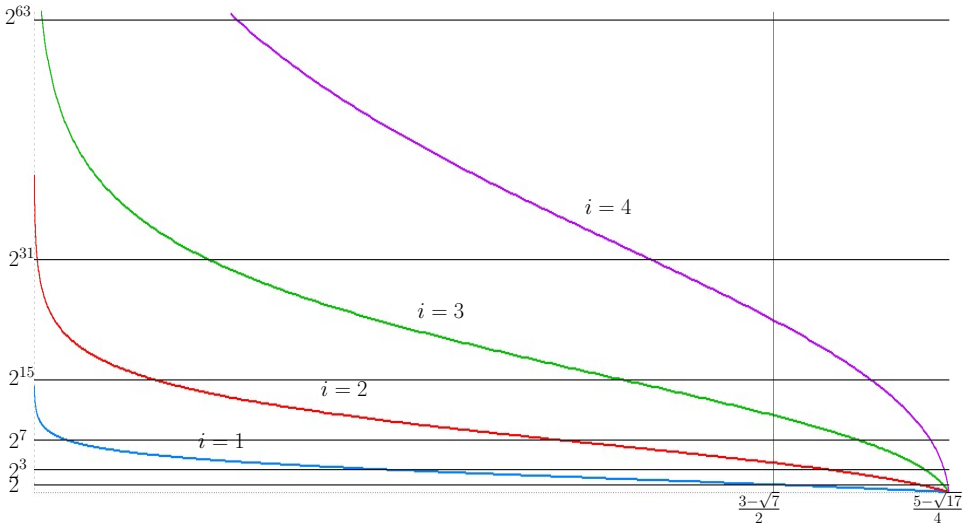

FIGURE 3. Values of $log_2(u_i/u_0)$ in function of u_0 for $i = 1, \ldots, 4$.

typically no more than 2^{24} or 2^{53}) times an exponent. (The IEEE-754 standard for computer arithmetic (The Institute of Electrical and Electronics Engineers Inc, 2008) is taught at elementary numerical analysis courses, see for instance Higham (2002, Ch.2)).

By using floating point numbers, a huge gain of speed is obtained with regard to exact representation of, say, algebraic numbers. However, computations are inexact (by a typical factor of 2^{-24} or 2^{-53}). Therefore, we need to consider **inexact** Newton iteration. An obvious modification of the proof of Theorem 4.2 gives us the following statement:

Theorem 4.9. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let ζ be a nondegenerate zero of $\mathbf{f}$. Let*

$$0 \leq 2\delta \leq u_0 \leq 2 - \frac{\sqrt{14}}{2} \simeq 0.129\cdots$$

Assume that

(1)
$$B = B\left(\zeta, \frac{u_0}{\gamma(\mathbf{f},\zeta)}\right) \subseteq \mathcal{D}.$$

(2) $\mathbf{x}_0 \in B$, *and the sequence* $\mathbf{x}_i$ *satisfies*
$$\|\mathbf{x}_{i+1} - N(\mathbf{f},\mathbf{x}_i)\|\gamma(\mathbf{f},\zeta) \leq \delta$$

(3) *The sequence* u_i *is defined inductively by*
$$u_{i+1} = \frac{u_i^2}{\psi(u_i)} + \delta.$$

Then the sequences u_i *and* $\mathbf{x}_i$ *are well-defined for all* i, $\mathbf{x}_i \in \mathcal{D}$, *and*
$$\frac{\|\mathbf{x}_i - \zeta\|}{\|\mathbf{x}_0 - \zeta\|} \leq \frac{u_i}{u_0} \leq \max\left(2^{-2^i+1}, 2\frac{\delta}{u_0}\right).$$

Proof. By hypothesis,
$$\frac{u_0}{\psi(u_0)} + \frac{\delta}{u_0} < 1$$
so the sequence u_i is decreasing and positive. For short, let $q = \frac{u_0}{\psi(u_0)} \leq 1/4$. By induction,
$$\frac{u_{i+1}}{u_0} \leq \frac{u_0}{\psi(u_i)}\left(\frac{u_i}{u_0}\right)^2 + \frac{\delta}{u_0} \leq \frac{1}{4}\left(\frac{u_i}{u_0}\right)^2 + \frac{\delta}{u_0}.$$

Assume that $u_i/u_0 \leq 2^{-2^i+1}$. In that case,
$$\frac{u_{i+1}}{u_0} \leq 2^{-2^{i+1}} + \frac{\delta}{u_0} \leq \max\left(2^{-2^{i+1}+1}, 2\frac{\delta}{u_0}\right).$$

Assume now that $2^{-2^i+1}, u_i/u_0 \leq 2\delta/u_0$. In that case,
$$\frac{u_{i+1}}{u_0} \leq \frac{\delta}{u_0}\left(\frac{\delta}{4u_0} + 1\right) \leq \frac{2\delta}{u_0} = \max\left(2^{-2^{i+1}+1}, 2\frac{\delta}{u_0}\right).$$

From now on we use the assumptions, notations and estimates of the proof of Theorem 4.2. Combining (5) and (8) in (6), we obtain again that
$$\|N(\mathbf{f},\mathbf{x})\| \leq \frac{\gamma\|\mathbf{x}\|^2}{\psi(\gamma\|\mathbf{x}\|)}.$$
This time, this means that
$$\|\mathbf{x}_{i+1}\|\gamma \leq \delta + \|N(\mathbf{f},\mathbf{x})\|\gamma \leq \delta + \frac{\gamma^2\|\mathbf{x}\|^2}{\psi(\gamma\|\mathbf{x}\|)}.$$

By induction that $\|\mathbf{x}_i - \zeta\|\gamma(\mathbf{f},\zeta) < u_i$ and we are done. □

Exercise 4.1. Consider the following series, defined in $\mathbb{C}^2$:

$$g(x) = \sum_{i=0}^{\infty} x_1^i x_2^i.$$

Compute its radius of convergence. What is its domain of absolute convergence ?

Exercise 4.2. The objective of this exercise is to produce a non-optimal algorithm to approximate $\sqrt{y}$. In order to do that, consider the mapping $f(x) = x^2 - y$.

(1) Compute $\gamma(f, x)$.
(2) Show that for $1 \leq y \leq 4$, $x_0 = 1/2 + y/2$ is an approximate zero of the first kind for x, associated to y.
(3) Write down an algorithm to approximate $\sqrt{y}$ up to relative accuracy 2^{-63}.

Exercise 4.3. Let $\mathbf{f}$ be an analytic map between Banach spaces, and assume that ζ is a nondegenerate zero of $\mathbf{f}$.

(1) Write down the Taylor series of $D\mathbf{f}(\zeta)^{-1}(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\zeta))$.
(2) Show that if $\mathbf{f}(\mathbf{x}) = 0$, then

$$\gamma(\mathbf{f}, \zeta)\|\mathbf{x} - \zeta\| \geq 1/2.$$

This shows that two nondegenerate zeros cannot be at a distance less than $1/2\gamma(\mathbf{f}, \zeta)$. Results of this type appeared in Dedieu (1997a), but some of them were known before Malajovich (1993, Th.16).

5. Estimates from data at a point

Theorem 4.2 guarantees quadratic convergence in a neighborhood of a known zero ζ. In practical situations, ζ is not known. A major result in alpha-theory is the criterion to detect an approximate zero with just local information. We need to slightly modify the definition.

Definition 5.1 (Approximate zero of the second kind)**.** Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be as above. An **approximate zero of the second kind** associated to $\zeta \in \mathcal{D}$, $\mathbf{f}(\zeta) = 0$, is a point $\mathbf{x}_0 \in \mathcal{D}$, such that

(1) The sequence $(\mathbf{x})_i$ defined inductively by $\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i)$ is well-defined (each $\mathbf{x}_i$ belongs to the domain of $\mathbf{f}$ and $D\mathbf{f}(\mathbf{x}_i)$ is invertible and bounded).
(2)

$$\|\mathbf{x}_{i+1} - \mathbf{x}_i\| \leq 2^{-2^i+1}\|\mathbf{x}_1 - \mathbf{x}_0\|.$$

(3) $\lim_{i\to\infty} \mathbf{x}_i = \zeta$.

For detecting approximate zeros of the second kind, we need:

Definition 5.2 (Smale's β and α invariants)**.**

$$\beta(\mathbf{f},\mathbf{x}) = \|D\mathbf{f}(\mathbf{x})^{-1}\mathbf{f}(\mathbf{x})\| \text{ and } \alpha(\mathbf{f},\mathbf{x}) = \beta(\mathbf{f},\mathbf{x})\gamma(\mathbf{f},\mathbf{x}).$$

The β invariant can be interpreted as the size of the Newton step $N(\mathbf{f},\mathbf{x}) - \mathbf{x}$.

Theorem 5.3 (Smale)**.** *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let*

$$\alpha \leq \alpha_0 = \frac{13 - 3\sqrt{17}}{4}.$$

Define

$$r_0 = \frac{1 + \alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha} \text{ and } r_1 = \frac{1 - 3\alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha}.$$

Let $\mathbf{x}_0 \in \mathcal{D}$ be such that $\alpha(\mathbf{f},\mathbf{x}_0) \leq \alpha$ and assume furthermore that $B(\mathbf{x}_0, r_0\beta(\mathbf{f},\mathbf{x}_0)) \subseteq \mathcal{D}$. Then,

1. *$\mathbf{x}_0$ is an approximate zero of the second kind, associated to some zero $\zeta \in \mathcal{D}$ of $\mathbf{f}$.*
2. *Moreover, $\|\mathbf{x}_0 - \zeta\| \leq r_0\beta(\mathbf{f},\mathbf{x}_0)$.*
3. *Let $\mathbf{x}_1 = N(\mathbf{f},\mathbf{x}_0)$. Then $\|\mathbf{x}_1 - \zeta\| \leq r_1\beta(\mathbf{f},\mathbf{x}_0)$.*

The constant α_0 is the largest possible with those properties.

This theorem appeared in Smale (1986). The value for α_0 was found by Wang Xinghua Wang Xinghua (1993). Numerically,

$$\alpha_0 = 0.157,670,780,786,754,587,633,942,608,019\cdots$$

Other useful numerical bounds, under the hypotheses of the theorem, are:

$$r_0 \leq 1.390,388,203\cdots \text{ and } r_1 \leq 0.390,388,203\cdots.$$

The proof of Theorem 5.3 follows from the same method as the one for Theorem 4.2. We first define the 'worst' real function with respect to Newton iteration. Let us fix $\beta, \gamma > 0$. Define

$$h_{\beta\gamma}(t) = \beta - t + \frac{\gamma t^2}{1 - \gamma t} = \beta - t + \gamma t^2 + \gamma^2 t^3 + \cdots.$$

We assume for the time being that $\alpha = \beta\gamma < 3 - 2\sqrt{2} = 0.1715\cdots$. This guarantees that $h_{\beta\gamma}$ has two distinct zeros $\zeta_1 = \frac{1+\alpha-\sqrt{\Delta}}{4\gamma}$ and $\zeta_2 = \frac{1+\alpha+\sqrt{\Delta}}{4\gamma}$ with of course $\Delta = (1+\alpha)^2 - 8\alpha$. An useful expression is the product formula

$$h_{\beta\gamma}(x) = 2\frac{(x - \zeta_1)(x - \zeta_2)}{\gamma^{-1} - x}. \tag{9}$$

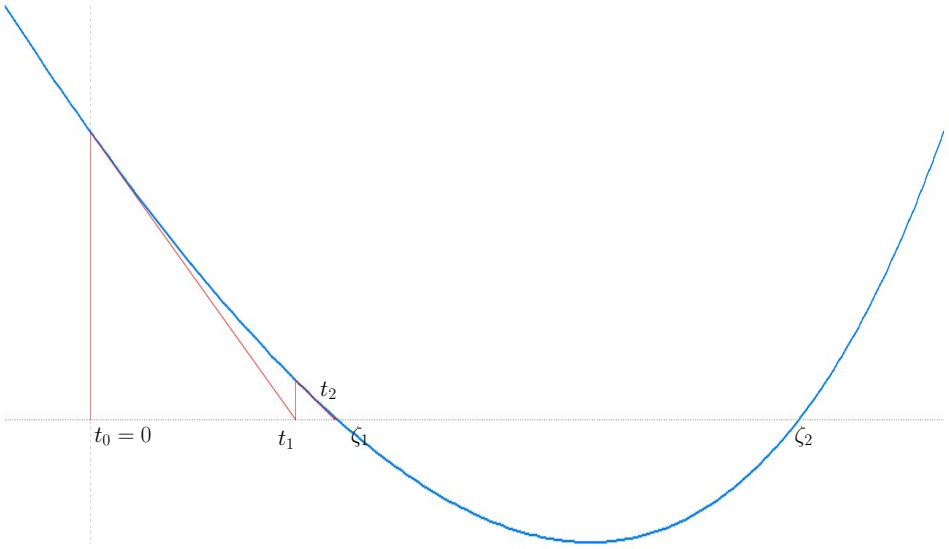

FIGURE 4. $y = h_{\beta\gamma}(t)$.

From (9), $h_{\beta\gamma}$ has also a pole at γ^{-1}. We have always $0 < \zeta_1 < \zeta_2 < \gamma^{-1}$.

The function $h_{\beta\gamma}$ is, among the functions with $h'(0) = -1$ and $\beta(h, 0) \leq \beta$ and $\gamma(h, 0) \leq \gamma$, the one that has the first zero ζ_1 furthest away from the origin.

Proposition 5.4. *Let $\beta, \gamma > 0$, with $\alpha = \beta\gamma \leq 3 - 2\sqrt{2}$. let $h_{\beta\gamma}$ be as above. Define recursively $t_0 = 0$ and $t_{i+1} = N(h_{\beta\gamma}, t_i)$. then*

$$t_i = \zeta_1 \frac{1 - q^{2^i - 1}}{1 - \eta q^{2^i - 1}}, \tag{10}$$

with

$$\eta = \frac{\zeta_1}{\zeta_2} = \frac{1 + \alpha - \sqrt{\Delta}}{1 + \alpha + \sqrt{\Delta}} \text{ and } q = \frac{\zeta_1 - \gamma\zeta_1\zeta_2}{\zeta_2 - \gamma\zeta_1\zeta_2} = \frac{1 - \alpha - \sqrt{\Delta}}{1 - \alpha + \sqrt{\Delta}}.$$

Proof. By differentiating (9), one obtains

$$h'_{\beta\gamma}(t) = h_{\beta\gamma}(t) \left(\frac{1}{t - \zeta_1} + \frac{1}{t - \zeta_2} + \frac{1}{\gamma^{-1} - t} \right)$$

and hence the Newton operator is

$$N(h_{\beta\gamma}, t) = t - \frac{1}{\frac{1}{t-\zeta_1} + \frac{1}{t-\zeta_2} + \frac{1}{\gamma^{-1}-t}}.$$

A tedious calculation shows that $N(h_{\beta\gamma}, t)$ is a rational function of degree 2. Hence, it is defined by 5 coefficients, or by 5 values.

In order to solve the recurrence for t_i, we change coordinates using a fractional linear transformation. As the Newton operator will have two attracting fixed points (ζ_1 and ζ_2), we will map those points to 0 and ∞ respectively. For convenience, we will map $t_0 = 0$ into $y_0 = 1$. Therefore, we set

$$S(t) = \frac{\zeta_2 t - \zeta_1\zeta_2}{\zeta_1 t - \zeta_1\zeta_2} \quad \text{and} \quad S^{-1}(y) = \frac{-\zeta_1\zeta_2 y + \zeta_1\zeta_2}{-\zeta_1 y + \zeta_2}$$

Let us look at the sequence $y_i = S(t_i)$. By construction $y_0 = 1$, and subsequent values are given by the recurrence

$$y_{i+1} = S(N(h_{\beta\gamma}, S^{-1}(y_i))).$$

It is an exercise to check that

$$y_{i+1} = q y_i^2, \tag{11}$$

Therefore we have $y_i = q^{2^i-1}$, and equation (10) holds. □

Proposition 5.5. *Under the conditions of Proposition 5.4, 0 is an approximate zero of the second kind for $h_{\beta\gamma}$ if and only if*

$$\alpha = \beta\gamma \leq \frac{13 - 3\sqrt{17}}{4}.$$

Proof. Using the closed form for t_i, we get:

$$\begin{aligned} t_{i+1} - t_i &= \frac{1 - q^{2^{i+1}-1}}{1 - \eta q^{2^{i+1}-1}} - \frac{1 - q^{2^i-1}}{1 - \eta q^{2^i-1}} \\ &= q^{2^i-1} \frac{(1-\eta)(1-q^{2^i})}{(1 - \eta q^{2^{i+1}-1})(1 - \eta q^{2^i-1})} \end{aligned}$$

In the particular case $i = 0$,

$$t_1 - t_0 = \frac{1-q}{1-\eta q} = \beta$$

Hence

$$\frac{t_{i+1} - t_i}{\beta} = C_i q^{2^i-1}$$

with

$$C_i = \frac{(1-\eta)(1-\eta q)(1-q^{2^i})}{(1-q)(1-\eta q^{2^{i+1}-1})(1-\eta q^{2^i-1})}.$$

Thus, $C_0 = 1$. The reader shall verify in Exercise 5.1 that C_i is a non-increasing sequence. Its limit is non-zero.

From the above, it is clear that 0 is an approximate zero of the second kind if and only if $q \leq 1/2$. Now, if we clear denominators and rearrange terms in $(1+\alpha-\sqrt{\Delta})/(1+\alpha+\sqrt{\Delta}) = 1/2$, we obtain the second degree polynomial

$$2\alpha^2 - 13\alpha + 2 = 0.$$

This has solutions $(13 \pm \sqrt{17})/2$. When $0 \leq \alpha \leq \alpha_0 = (13-\sqrt{17})/2$, the polynomial values are positive and hence $q \leq 1/2$. □

Proof of Th.5.3. Let $\beta = \beta(\mathbf{f}, \mathbf{x}_0)$ and $\gamma = \gamma(\mathbf{f}, \mathbf{x}_0)$. Let $h_{\beta\gamma}$ and the sequence t_i be as in Proposition 5.4. By construction, $\|\mathbf{x}_1 - \mathbf{x}_0\| = \beta = t_1 - t_0$. We use the following notations:

$$\beta_i = \beta(\mathbf{f}, \mathbf{x}_i) \text{ and } \gamma_i = \gamma(\mathbf{f}, \mathbf{x}_i).$$

Those will be compared to

$$\hat{\beta}_i = \beta(h_{\beta\gamma}, t_i)) \text{ and } \hat{\gamma}_i = \gamma(h_{\beta\gamma}, t_i)).$$

Induction hypothesis: $\beta_i \leq \hat{\beta}_i$ and for all $l \geq 2$,

$$\|D\mathbf{f}(\mathbf{x}_i)^{-1} D^l \mathbf{f}(\mathbf{x}_i)\| \leq -\frac{h^{(l)}_{\beta\gamma}(t_i)}{h'_{\beta\gamma}(t_i)}.$$

The initial case when $i = 0$ holds by construction. So let us assume that the hypothesis holds for i. We will estimate

$$\beta_{i+1} \leq \|D\mathbf{f}(\mathbf{x}_{i+1})^{-1} D\mathbf{f}(\mathbf{x}_i)\| \|D\mathbf{f}(\mathbf{x}_i)^{-1}\mathbf{f}(\mathbf{x}_{i+1})\| \tag{12}$$

and

$$\gamma_{i+1} \leq \|D\mathbf{f}(\mathbf{x}_{i+1})^{-1} D\mathbf{f}(\mathbf{x}_i)\| \frac{\|D\mathbf{f}(\mathbf{x}_i)^{-1} D^k \mathbf{f}(\mathbf{x}_{i+1})\|}{k!}. \tag{13}$$

By construction, $\mathbf{f}(\mathbf{x}_i) + D\mathbf{f}(\mathbf{x}_i)(\mathbf{x}_{i+1} - \mathbf{x}_i) = 0$. The Taylor expansion of $\mathbf{f}$ at $\mathbf{x}_i$ is therefore

$$D\mathbf{f}(\mathbf{x}_i)^{-1}\mathbf{f}(\mathbf{x}_{i+1}) = \sum_{k \geq 2} \frac{D\mathbf{f}(\mathbf{x}_i)^{-1} D^k \mathbf{f}(\mathbf{x}_i)(\mathbf{x}_{i+1} - \mathbf{x}_i)^k}{k!}$$

Passing to norms,

$$\|D\mathbf{f}(\mathbf{x}_i)^{-1}\mathbf{f}(\mathbf{x}_{i+1})\| \leq \frac{\beta_i^2 \gamma_i}{1 - \gamma_i}$$

The same argument shows that

$$-\frac{h_{\beta\gamma}(t_{i+1})}{h'_{\beta\gamma}(t_i)} = \frac{\beta(h_{\beta\gamma}, t_i)^2 \gamma(h_{\beta\gamma}, t_i)}{1 - \gamma(h_{\beta\gamma}, t_i)}$$

From Lemma 4.6,

$$\|D\mathbf{f}(\mathbf{x}_{i+1})^{-1}D\mathbf{f}(\mathbf{x}_i)\| \leq \frac{(1-\beta_i\gamma_i)^2}{\psi(\beta_i\gamma_i)}.$$

Also, computing directly,

$$\frac{h'_{\beta\gamma}(t_{i+1})}{h'_{\beta\gamma}(t_i)} = \frac{(1-\hat\beta\hat\gamma)^2}{\psi(\hat\beta\hat\gamma)}. \tag{14}$$

We established that

$$\beta_{i+1} \leq \frac{\beta_i^2\gamma_i(1-\beta_i\gamma_i)}{\psi(\beta_i\gamma_i)} \leq \frac{\hat\beta_i^2\hat\gamma_i(1-\hat\beta_i\hat\gamma_i)}{\psi(\hat\beta_i\hat\gamma_i)} = \hat\beta_{i+1}.$$

Now the second part of the induction hypothesis:

$$D\mathbf{f}(\mathbf{x}_i)^{-1}D^l\mathbf{f}(\mathbf{x}_{i+1}) = \sum_{k\geq 0}\frac{1}{k!}\frac{D\mathbf{f}(\mathbf{x}_i)^{-1}D^{k+l}\mathbf{f}(\mathbf{x}_i)(\mathbf{x}_{i+1}-\mathbf{x}_i)^k}{k+l}$$

Passing to norms and invoking the induction hypothesis,

$$\|D\mathbf{f}(\mathbf{x}_i)^{-1}D^l\mathbf{f}(\mathbf{x}_{i+1})\| \leq \sum_{k\geq 0} -\frac{h^{(k+l)}_{\beta\gamma}(t_i)\hat\beta_i^k}{k!h'_{\beta\gamma}(t_i)}$$

and then using Lemma 4.6 and (14),

$$\|D\mathbf{f}(\mathbf{x}_{i+1})^{-1}D^l\mathbf{f}(\mathbf{x}_{i+1})\| \leq \frac{(1-\hat\beta_i\hat\gamma_i)^2}{\psi(\hat\beta_i\hat\gamma_i)}\sum_{k\geq 0} -\frac{h^{(k+l)}_{\beta\gamma}(t_i)\hat\beta_i^k}{k!h'_{\beta\gamma}(t_i)}.$$

A direct computation similar to (14) shows that

$$-\frac{h^{(k+l)}_{\beta\gamma}(t_{i+1})}{k!h'_{\beta\gamma}(t_{i+1})} = \frac{(1-\hat\beta_i\hat\gamma_i)^2}{\psi(\hat\beta_i\hat\gamma_i)}\sum_{k\geq 0} -\frac{h^{(k+l)}_{\beta\gamma}(t_i)\hat\beta_i^k}{k!h'_{\beta\gamma}(t_i)}.$$

and since the right-hand-terms of the last two equations are equal, the second part of the induction hypothesis proceeds. Dividing by $l!$, taking $l-1$-th roots and maximizing over all l, we deduce that $\gamma_i \leq \hat\gamma_i$.

Proposition 5.5 then implies that $\mathbf{x}_0$ is an approximate zero.

The second and third statement follow respectively from

$$\|\mathbf{x}_0 - \zeta\| \leq \beta_0 + \beta_1 + \cdots = \zeta_1$$

and

$$\|\mathbf{x}_1 - \zeta\| \leq \beta_1 + \beta_2 + \cdots = \zeta_1 - \beta.$$

□

	1/32	1/16	1/10	1/8	$\frac{13-3\sqrt{17}}{4}$
1	4.854	3.683	2.744	2.189	1.357
2	14.472	10.865	7.945	6.227	3.767
3	33.700	25.195	18.220	14.41	7.874
4	72.157	53.854	38.767	29.648	15.881
5	149.71	111.173	79.861	60.864	31.881
6	302.899	225.811	162.49	123.295	63.881

TABLE 2. Values of $-log_2(\|\mathbf{x}_i - \zeta\|/\beta)$ in function of α and i.

The same issues as in Theorem 4.2 arise. First of all, we actually proved a sharper statement. Namely,

Theorem 5.6. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let*

$$\alpha \leq 3 - 2\sqrt{2}.$$

Define

$$r = \frac{1 + \alpha - \sqrt{1 - 6\alpha + \alpha^2}}{4\alpha}.$$

Let $\mathbf{x}_0 \in \mathcal{D}$ be such that $\alpha(\mathbf{f}, \mathbf{x}_0) \leq \alpha$ and assume furthermore that $B(\mathbf{x}_0, r\beta(\mathbf{f}, \mathbf{x}_0)) \subseteq \mathcal{D}$. Then, the sequence $\mathbf{x}_{i+1} = N(\mathbf{f}, \mathbf{x}_i)$ is well defined, and there is a zero $\zeta \in \mathcal{D}$ of $\mathbf{f}$ such that

$$\|\mathbf{x}_i - \zeta\| \leq q^{2^i - 1} \frac{1 - \eta}{1 - \eta q^{2^i - 1}} r\beta(\mathbf{f}, \mathbf{x}_0).$$

for η and q as in Proposition 5.4.

Table 2 and Figure 5 show how fast $\|\mathbf{x}_i - \zeta\|/\beta$ decreases in terms of α and i.

The final issue is robustness. There is no obvious modification of the proof of Theorem 5.3 to provide a nice statement, so we will rely on Theorem 4.9 indeed.

Theorem 5.7. *Let $\mathbf{f} : \mathcal{D} \subseteq \mathbb{E} \to \mathbb{F}$ be an analytic map between Banach spaces. Let δ, α and $u0$ satisfy*

$$0 \leq 2\delta < u_0 = \frac{r\alpha}{(1 - r\alpha)\psi(r\alpha)} < 2 - \frac{\sqrt{14}}{2}$$

with $r = \frac{1+\alpha-\sqrt{1-6\alpha+\alpha^2}}{4\alpha}$. Assume that

(1)
$$B = B\left(\mathbf{x}_0, 2r\beta(\mathbf{f}, \mathbf{x}_0)\right) \subseteq \mathcal{D}.$$

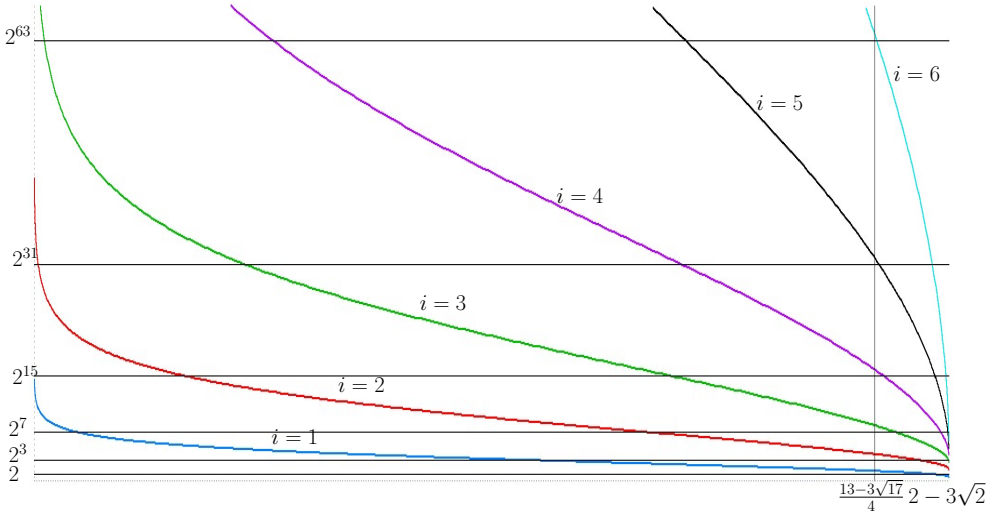

FIGURE 5. Values of $-log_2(\|\mathbf{x}_i - \zeta\|/\beta)$ in function of α for $i = 1$ to 6.

(2) $\mathbf{x}_0 \in B$, *and the sequence* $\mathbf{x}_i$ *satisfies*

$$\|\mathbf{x}_{i+1} - N(\mathbf{f}, \mathbf{x}_i)\| \frac{r\beta(f, x_0)}{(1 - r\alpha)\psi(r\alpha)} \leq \delta$$

(3) *The sequence* u_i *is defined inductively by*

$$u_{i+1} = \frac{u_i^2}{\psi(u_i)} + \delta.$$

Then the sequences u_i *and* $\mathbf{x}_i$ *are well-defined for all* i, $\mathbf{x}_i \in \mathcal{D}$, *and*

$$\frac{\|\mathbf{x}_i - \zeta\|}{\|\mathbf{x}_1 - \mathbf{x}_0\|} \leq \frac{ru_i}{u_0} \leq r \max\left(2^{-2^i+1}, 2\frac{\delta}{u_0}\right).$$

Numerically, $\alpha_0 = 0.074,290\cdots$ satisfies the hypothesis of the Theorem. A version of this theorem (not as sharp, and another metric) appeared as Theorem 2 in Malajovich (1994).

The following Lemma will be useful:

Lemma 5.8. *Assume that* $u = \gamma(\mathbf{f}, \mathbf{x})\|\mathbf{x} - \mathbf{y}\| \leq 1 - \sqrt{2}/2$. *Then,*

$$\gamma(\mathbf{f}, \mathbf{y}) \leq \frac{\gamma(\mathbf{f}, \mathbf{x})}{(1 - u)\psi(u)}.$$

Proof. In order to estimate the higher derivatives, we expand:

$$\frac{1}{l!} D\mathbf{f}(\mathbf{x})^{-1} D^l \mathbf{f}(\mathbf{y}) = \sum_{k \geq 0} \binom{k+l}{l} \frac{D\mathbf{f}(\mathbf{x})^{-1} D^{k+l}\mathbf{f}(\mathbf{x})(\mathbf{y} - \mathbf{x})^k}{k + l}$$

and by Lemma 4.3 for $d = l + 1$,

$$\frac{1}{l!}\|D\mathbf{f}(\mathbf{x})^{-1}D^l\mathbf{f}(\mathbf{y})\| \leq \frac{\gamma(\mathbf{f},\mathbf{x})^{l-1}}{(1-u)^{l+1}}.$$

Combining with Lemma 4.6,

$$\frac{1}{l!}\|D\mathbf{f}(\mathbf{y})^{-1}D^l\mathbf{f}(\mathbf{y})\| \leq \frac{\gamma(\mathbf{f},\mathbf{x})^{l-1}}{(1-u)^{l-1}\psi(u)}.$$

Taking the $l-1$-th power,

$$\gamma(\mathbf{f},\mathbf{y}) \leq \frac{\gamma(\mathbf{f},\mathbf{x})}{(1-u)\psi(u)}.$$

□

Proof of Theorem 5.7. We have necessarily $\alpha < 3 - 2\sqrt{2}$ or r is undefined. Then (Theorem 5.6) there is a zero ζ of $\mathbf{f}$ with $\|\mathbf{x}_0 - \zeta\| \leq r\beta(f, x_0)$. Then, Lemma 5.8 implies that $\|\mathbf{x}_0 - \zeta\|\gamma(\mathbf{f},\zeta) \leq u_0$. Now apply Theorem 4.9.

□

Exercise 5.1. The objective of this exercise is to show that C_i is non-increasing.

(1) Show the following trivial lemma: **If** $0 \leq s < a \leq b$**, then** $\frac{a-s}{b-s} \leq \frac{a}{b}$.
(2) Deduce that $q \leq \eta$.
(3) Prove that $C_{i+1}/C_i \leq 1$.

Exercise 5.2. Show that

$$\zeta_1\gamma(\zeta_1) = \frac{1+\alpha-\sqrt{\Delta}}{3-\alpha+\sqrt{\Delta}}\frac{1}{\psi\left(\frac{1+\alpha-\sqrt{\Delta}}{4}\right)}.$$

Part 2. Inclusion and exclusion

6. Eckart-Young theorem

The following classical theorem in linear algebra is known as the **singular value decomposition** (**svd** for short).

Theorem 6.1. *Let $A : \mathbb{R}^n \mapsto \mathbb{R}^m$ (resp. $\mathbb{C}^n \to \mathbb{C}^m$) be linear. Then, there are $\sigma_1 \geq \cdots \geq \sigma_r > 0$, $r \leq m, n$, such that*

$$A = U\Sigma V^*$$

with $U \in O(m)$ (resp. $U(m)$), $V \in O(n)$ (resp. $U(n)$) and $\Sigma_{ij} = \sigma_i$ for $i = j \leq r$ and 0 otherwise.

It is due to Sylvester (real $n \times n$ matrices) and to Eckart and Young (1939) in the general case, now exercise 6.1 below.

Σ is a $m \times n$ matrix. It is possible to rewrite this in an 'economical' formulation with Σ an $r \times r$ matrix, U and V orthogonal (resp. unitary) $m \times r$ and $n \times r$ matrices. The numbers $\sigma_1, \ldots, \sigma_r$ are called **singular values** of A. They may be computed by extracting the positive square root of the non-zero eigenvalues of A^*A or AA^*, whatever matrix is smaller. The operator and Frobenius norm of A may be written in terms of the σ_i's:

$$\|A\|_2 = \sigma_1 \qquad \|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}.$$

The discussion and the results above hold when A is a linear operator between finite dimensional inner product spaces. It suffices to choose an orthonormal basis, and apply Theorem 6.1 to the corresponding matrix.

When $m = n = r$, $\|A^{-1}\|_2 = \sigma_n$. In this case, the **condition number** of A for linear solving is defined as

$$\kappa(A) = \|A\|_* \|A^{-1}\|_{**}.$$

The choice of norms is arbitrary, as long as operator and vector norms are consistent. Two canonical choices are

$$\kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 \text{ and } \kappa_D(A) = \|A\|_F \|A^{-1}\|_2.$$

The second choice was suggested by Demmel (1988). Using that definition he obtained bounds on the probability that a matrix is poorly conditioned. The exact probability distribution for the most usual probability measures in matrix space was computed in Edelman (1992).

Assume that $A(t)\mathbf{x}(t) \equiv \mathbf{b}(t)$ is a family of problems and solutions depending smoothly on a parameter t. Differentiating implicitly,

$$\dot{A}\mathbf{x} + A\dot{\mathbf{x}} = \dot{\mathbf{b}}$$

which amounts to

$$\dot{\mathbf{x}} = A^{-1}\dot{\mathbf{b}} - A^{-1}\dot{A}\mathbf{x}.$$

Passing to norms and to relative errors, we quickly obtain

$$\frac{\|\dot{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \kappa_D(A) \left(\frac{\|\dot{A}\|_F}{\|A\|_F} + \frac{\|\dot{\mathbf{b}}\|}{\|\mathbf{b}\|} \right).$$

This bounds the relative error in the solution $\mathbf{x}$ in terms of the relative error in the coefficients. The usual paradigm in numerical linear

algebra dates from Turing (1948) and Wilkinson (1994). After the rounding-off during computation, we obtain the exact solution of a perturbed system. Bounds for the perturbation or **backward error** are found through line by line analysis of the algorithm. The output error or **forward error** is bounded by the backward error, times the condition number.

Condition numbers provide therefore an important metric invariant for numerical analysis problems. A geometric interpretation in the case of linear equation solving is:

Theorem 6.2. *Let A be a nondegenerate square matrix.*

$$\|A^{-1}\|_2 = \min_{\det(A+B)=0} \|B\|_F$$

In particular, this implies that

$$\kappa_D(A)^{-1} = \min_{\det(A+B)=0} \frac{\|B\|_F}{\|A\|_F}$$

A pervading principle in the subject is: **the inverse of the condition number is related to the distance to the ill-posed problems**.

It is possible to define the condition number for a full-rank non-square matrix by

$$\kappa_D(A) = \|A\|_F \ \sigma_{\min(m,n)}(A)^{-1}.$$

Theorem 6.3. (Eckart and Young, 1936) *Let A be an $m \times n$ matrix of rank r. Then,*

$$\sigma_r(A)^{-1} = \min_{\sigma_r(A+B)=0} \|B\|_F.$$

In particular, if $r = \min(m, n)$,

$$\kappa_D(A)^{-1} = \min_{\sigma_r(A+B)=0} \frac{\|B\|_F}{\|A\|_F}.$$

Exercise 6.1. Prove Theorem 6.1. Hint: let u, v, σ such that $Av = \sigma u$ with σ maximal, $\|u\| = 1$, $\|v\| = 1$. What can you say about $A_{|v^\perp}$?

Exercise 6.2. Prove Theorem 6.3.

Exercise 6.3. Assume furthermore that $m < n$. Show that the same interpretation for the condition number still holds, namely the norm of the perturbation of **some** solution is bounded by the condition number, times the perturbation of the input.

7. The space of homogeneous polynomial systems

We will denote by $\mathcal{H}_d^{\mathbb{R}}$ the space of polynomials of degree d in $n+1$ variables. This space can be assimilated to the space of symmetric d-linear forms. For instance, when $d=2$, the polynomial

$$f(x_0,x_1) = f_0x_0^2 + f_1x_0x_1 + f_2x_1^2 = \begin{bmatrix} x_0 & x_1 \end{bmatrix} \begin{bmatrix} f_0 & f_1/2 \\ f_1/2 & f_0 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix}$$

can be assimilated to a symmetric bilinear form and can be represented by a matrix. In general, a homogeneous polynomial can be represented by a symmetric tensor

$$f(\mathbf{x}) = \sum_{|\mathbf{a}|=d} f_{\mathbf{a}} x_0^{a_0} \cdots x_n^{a_n} = \sum_{0\le i_1,\dots,i_d\le n} T_{i_1 i_2 \dots i_d} x_{i_1} x_{i_2} \cdots x_{i_d}$$

where

$$f_{\mathbf{a}} = \sum_{\mathbf{a}=\mathrm{e}_{i_1}+\mathrm{e}_{i_2}+\cdots \mathrm{e}_{i_d}} T_{i_1 i_2 \dots i_d}.$$

The canonical inner product for tensors is given by

$$\langle S, T\rangle = \sum_{0\le i_1,\dots,i_d\le n} S_{i_1 i_2 \dots i_d} T_{i_1 i_2 \dots i_d}$$

The same inner product for polynomials is written

$$\langle f, g\rangle = \sum_{|\mathbf{a}|=d} \frac{f_{\mathbf{a}} g_{\mathbf{a}}}{\binom{d}{\mathbf{a}}}.$$

where $\binom{d}{\mathbf{a}} = \frac{d!}{a_0! a_1! \cdots a_n!}$ is the coefficient of $(x_0+\cdots+x_n)^d$ in x^a.

Lemma 7.1. *Let Q be an orthogonal $n\times n$ matrix, that is $Q^TQ = I$. Then,*

$$\langle f\circ Q, g\circ Q\rangle = \langle f, g\rangle$$

Exercise 7.1. Prove Lemma 7.1

We say that the above inner product is **invariant under orthogonal action**. We will always assume this inner-product for $\mathcal{H}_d^{\mathbb{R}}$.

It is also important to notice that $\mathcal{H}_d^{\mathbb{R}}$ is that it is a **reproducing kernel space**. Let

$$K_d(\mathbf{x},\mathbf{y}) = \langle \mathbf{x}, \mathbf{y}\rangle^d.$$

Then

$$f(\mathbf{y}) = \langle f(\cdot), K_d(\cdot, \mathbf{y})\rangle,$$
$$Df(\mathbf{y})\mathbf{u} = \langle f(\cdot), D_{\mathbf{y}} K_d(\cdot,\mathbf{y})\mathbf{u}\rangle,$$

etc...

8. The condition number

Now, let's denote by $\mathcal{H}_{\mathbf{d}}^{\mathbb{R}}$ the space of systems of homogeneous polynomials of degree $\mathbf{d} = (d_1, \ldots, d_n)$. The **condition number** measures how does the solution of an equation depends upon the coefficients.

Therefore, assume that both a polynomial system $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ and a point $\mathbf{x} \in S(\mathbb{R}^{n+1})$ depend upon a parameter t. Say,

$$\mathbf{f}_t(\mathbf{x}_t) \equiv 0.$$

Differentiating, one gets

$$D\mathbf{f}_t(\mathbf{x}_t)\dot{\mathbf{x}}_t = -\dot{\mathbf{f}}_t(\mathbf{x}_t)$$

so

$$\|\dot{\mathbf{x}}_t\| \leq \|D\mathbf{f}_t(\mathbf{x}_t)^{-1}_{|\mathbf{x}_t^\perp}\| \|\dot{\mathbf{f}}_t(\mathbf{x}_t)\|. \tag{15}$$

The **normalized condition number** is defined for $\mathbf{f} \in \mathcal{H}_{\mathbf{d}}^{\mathbb{R}}$ and $\mathbf{x} \in \mathbb{R}^{n+1}$ as

$$\mu(\mathbf{f}, \mathbf{x}) = \|\mathbf{f}\| \left\| \left(\begin{bmatrix} d_1^{-1/2}\|\mathbf{x}\|^{-d_1+1} & & \\ & \ddots & \\ & & d_n^{-1/2}\|\mathbf{x}\|^{-d_n+1} \end{bmatrix} D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp} \right)^{-1} \right\|.$$

In the special case $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ and $\mathbf{x} \in S(\mathbb{R}^{n+1})$,

$$\mu(\mathbf{f}, \mathbf{x}) = \left\| \left(\begin{bmatrix} d_1^{-1/2} & & \\ & \ddots & \\ & & d_n^{-1/2} \end{bmatrix} D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp} \right)^{-1} \right\|.$$

Proposition 8.1.

(1) *If $\mathbf{f}_t$ and $\mathbf{x}_t$ are paths in $S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ and $S(\mathbb{R}^{n+1})$ respectively, and $\mathbf{f}_t(\mathbf{x}_t) \equiv 0$ then*

$$\|\dot{\mathbf{x}}_t\| \leq \mu(\mathbf{f}_t, \mathbf{x}_t)\|\dot{\mathbf{f}}_\mathbf{t}\|.$$

(2) *Let $\mathbf{x} \in S(\mathbb{R}^{n+1})$ be fixed. Then the mapping*

$$\begin{array}{rccl} \pi: & \mathcal{H}_{\mathbf{d}}^{\mathbb{R}} & \to & L(\mathbf{x}^\perp, \mathbb{R}^n), \\ & \mathbf{f} & \mapsto & \begin{bmatrix} d_1^{-1/2} & & & \\ & d_2^{-1/2} & & \\ & & \ddots & \\ & & & d_n^{-1/2} \end{bmatrix} D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp} \end{array}$$

restricts to an isometry $\pi_{|(\ker\pi)^\perp} : (\ker\pi)^\perp \to L(\mathbf{x}^\perp, \mathbb{R}^n)$.

(3) *Let* $\mathbf{f} \in S(\mathcal{H}^{\mathbb{R}}_{\mathbf{d}})$ *and* $\mathbf{x} \in S(\mathbb{R}^{n+1})$. *Then,*

$$\mu(\mathbf{f}, \mathbf{x}) = \frac{1}{\min\{\|\mathbf{f}-\mathbf{g}\| : D\mathbf{g}(\mathbf{x})_{|\mathbf{x}^\perp} \text{ singular}\}}.$$

(4) *If furthermore* $\mathbf{f}(\mathbf{x}) = 0$,

$$\mu(\mathbf{f}, \mathbf{x}) = \frac{1}{\min\{\|\mathbf{f}-\mathbf{g}\| : \mathbf{g}(\mathbf{x}) = 0 \text{ and } D\mathbf{g}(\mathbf{x})_{|\mathbf{x}^\perp} \text{ singular}\}}.$$

Proof. Item 1 follows from (15). In order to prove item 2, let $\mathbf{x} \in S(\mathbb{R}^{n+1})$ be fixed and let $\mathbf{f} \in \mathcal{H}^{\mathbb{R}}_{\mathbf{d}}$. Assume that $\mathbf{y} \perp \mathbf{x}$. We can write $\mathbf{f}(\mathbf{x}+\mathbf{y})$ as

$$\mathbf{f}(\mathbf{x}+\mathbf{y}) = \mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp}\mathbf{y} + \frac{1}{2}D^2\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp}(\mathbf{y}-\mathbf{x}, \mathbf{y}-\mathbf{x}) + \cdots$$

This suggests a decomposition of $\mathcal{H}^{\mathbb{R}}_{\mathbf{d}}$ into terms that are 'constant', 'linear' or 'higher order' at x.

$$\mathcal{H}^{\mathbb{R}}_{\mathbf{d}} = H_0 \oplus H_1 \oplus H_2 \oplus \cdots .$$

An orthonormal basis for H_1 would be

$$\left(\frac{1}{\sqrt{d}}\frac{\partial K_{d_i}(\cdot, \mathbf{x})}{\partial \mathbf{u}_j}\mathrm{e}_i\right)$$

where $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ is an orthonormal basis of $\mathbf{x}^\perp$ and $(\mathrm{e}_1, \ldots, \mathrm{e}_n)$ is the canonical basis of $\mathbb{R}^n$.

In this basis, the projection of $\mathbf{f}$ in H_1 is just

$$\begin{bmatrix} & \vdots & \\ \cdots & \left\langle \mathbf{f}_i, \frac{1}{\sqrt{d}}\frac{\partial K_{d_i}(\cdot,\mathbf{x})}{\partial \mathbf{u}_j}\right\rangle & \cdots \\ & \vdots & \end{bmatrix} = \begin{bmatrix} d_1^{-1/2} & & \\ & \cdots & \\ & & d_n^{-1/2} \end{bmatrix} D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp}.$$

Thus, the subspace H_1 of $\mathcal{H}^{\mathbb{R}}_{\mathbf{d}}$ is isomorphic to the space of $n \times n$ matrices. Moreover, $\pi : \mathcal{H}^{\mathbb{R}}_{\mathbf{d}} \to H_1$ is an orthogonal projection. Items 3 and 4 follow now easily from Theorem 6.3. □

Exercise 8.1. Deduce that for all $\mathbf{f} \in \mathcal{H}^{\mathbb{R}}_{\mathbf{d}}$, $0 \neq \mathbf{x} \in \mathbb{R}^{n+1}$, $\mu(\mathbf{f}, \mathbf{x}) \geq \sqrt{n}$.

We denote by $\rho(\mathbf{x}, \mathbf{y}) = \widehat{(\mathbf{x}0\mathbf{y})}$ the angular distance between $\mathbf{x} \in S^n$ and $\mathbf{y} \in S^n$. The following estimate is quite useful:

Theorem 8.2. *Let* $\mathbf{f}, \mathbf{g} \in S(\mathcal{H}^{\mathbb{R}}_{\mathbf{d}})$ *and let* $\mathbf{x}, \mathbf{y} \in S(\mathbb{R}^{n+1})$. *Let*

$$u = (\max d_i)\mu(\mathbf{f}, \mathbf{g})\rho(\mathbf{x}, \mathbf{y}) \text{ and } v = \mu(\mathbf{f}, \mathbf{x})\|\mathbf{f}-\mathbf{g}\|.$$

Then,

$$\frac{1}{1+u+v}\mu(\mathbf{f},\mathbf{x}) \leq \mu(\mathbf{g},\mathbf{y}) \leq \frac{1}{1-u-v}\mu(\mathbf{f},\mathbf{x}).$$

Remark 8.3. Similar formulas appeared in Bürgisser and Cucker (2011) and Dedieu et al. (2012). The final form here appeared in Malajovich (2011) and generalizes to the sparse condition number.

Proof. Let R be a rotation taking $\mathbf{y}$ to $\mathbf{x}$. Then, $\mu(\mathbf{g},\mathbf{y}) = \mu(\mathbf{g}\circ R,\mathbf{x})$. Moreover, it is easy to check that $\|\mathbf{g}\circ R - \mathbf{g}\| \leq (\max d_i)\rho(\mathbf{x},\mathbf{y})$. Thus,

$$\mu(\mathbf{f},\mathbf{x})\|\mathbf{f}-\mathbf{g}\circ R\| \leq (u+v).$$

Now, notice that Proposition 8.1(3) implies:

$$\frac{1}{\mu(\mathbf{f},\mathbf{x})} - \|\mathbf{f}-\mathbf{g}\circ R\|. \leq \frac{1}{\mu(\mathbf{g}\circ R,\mathbf{x})} \leq \frac{1}{\mu(\mathbf{f},\mathbf{x})} + \|\mathbf{f}-\mathbf{g}\circ R\|.$$

The theorem follows by taking inverses. □

9. The inclusion theorem

For any $\mathbf{x} \in S(\mathcal{H}^{\mathbb{R}}_{\mathbf{d}})$, we denote by $A_{\mathbf{x}}$ be the affine space $\mathbf{x}+\mathbf{x}^{\perp}$ and by $F_{\mathbf{x}} : A_{\mathbf{x}} \to \mathbb{R}^n$, $\mathbf{X} \mapsto \mathbf{f}(\mathbf{x}+\mathbf{X})$ the restriction of $\mathbf{f}$ to $A_{\mathbf{x}}$. Then $F_{\mathbf{x}}$ is an n-variate polynomial system of degree $\mathbf{d}$.

Lemma 9.1. (Shub and Smale, 1993)

$$\gamma(\mathbf{F}_{\mathbf{x}},0) \leq \frac{(\max d_i)^{3/2}}{2}\|\mathbf{f}\|\mu(\mathbf{f},\mathbf{x})$$

Proof. For simplicity assume $\|\mathbf{f}\| = 1$. Let $k \geq 2$ and

$$\Delta = \begin{bmatrix} \sqrt{d_1} & & \\ & \ddots & \\ & & \sqrt{d_n} \end{bmatrix}.$$

$$\begin{aligned} \frac{1}{k!}\left\|D\mathbf{F}_{\mathbf{x}}(0)^{-1}D^k\mathbf{F}_{\mathbf{x}}(0)\right\| &= \frac{1}{k!}\left\|D\mathbf{f}(\mathbf{x})^{-1}_{|\mathbf{x}^{\perp}}D^k\mathbf{f}(\mathbf{x})_{|\mathbf{x}^{\perp}}\right\| \\ &\leq \frac{1}{k!}\left\|D\mathbf{f}(\mathbf{x})^{-1}_{|\mathbf{x}^{\perp}}\Delta\right\|\left\|\Delta^{-1}D^k\mathbf{f}(\mathbf{x})_{|\mathbf{x}^{\perp}}\right\| \\ &\leq \mu(\mathbf{f},\mathbf{x})\ \frac{1}{k!}\left\|\Delta^{-1}D^k\mathbf{f}(\mathbf{x})_{|\mathbf{x}^{\perp}}\right\| \end{aligned}$$

Now, notice that

$$|D^k\mathbf{f}_i(\mathbf{x})| = |\langle \mathbf{f}_i, D^k K_{d_i}(\cdot,\mathbf{x})\rangle| \leq$$
$$\leq \|\mathbf{f}_i\| \sup_{\substack{\|\mathbf{u}_1\|=\cdots=\|\mathbf{u}_k\|=1 \\ \mathbf{u}_1,\ldots,\mathbf{u}_k\perp\mathbf{x}}} \|D^k K_{d_i}(\cdot,\mathbf{x})(\mathbf{u}_1,\ldots,\mathbf{u}_k)\|$$

where $K_{d_i}(\mathbf{y},\mathbf{x}) = \langle \mathbf{y},\mathbf{x}\rangle^{d_i}$ is the reproducing kernel of $\mathcal{H}^{\mathbb{R}}_{d_i}$. Differentiating K_{d_i} with respect to $\mathbf{y}$, one obtains:

$$\frac{1}{k!}D^k K_{d_i}(\mathbf{y},\mathbf{x})(\mathbf{u}_1,\dots,\mathbf{u}_k) = \binom{d_i}{k}\langle \mathbf{y},\mathbf{x}\rangle^{d-k}\langle y,\mathbf{u}_1\rangle\cdots\langle y,\mathbf{u}_k\rangle.$$

The norm of $\frac{1}{k!}D^k K_{d_i}(\mathbf{y},\mathbf{x})(\mathbf{u}_1,\dots,\mathbf{u}_k)$ (as a polynomial of $\mathbf{y}$) can be computed using the reproducing kernel property.

$$\begin{aligned}
\left\|\frac{1}{k!}D^k K_{d_i}(\cdot,\mathbf{x})(\mathbf{u}_1,\dots,\mathbf{u}_k)\right\|^2 = \\
&= \left\langle \frac{1}{k!}D^k K_{d_i}(\cdot,\mathbf{x})(\mathbf{u}_1,\dots,\mathbf{u}_k), \frac{1}{k!}D^k K_{d_i}(\cdot,\mathbf{x})(\mathbf{u}_1,\dots,\mathbf{u}_k)\right\rangle \\
&= \frac{1}{k!}\frac{\partial \mathbf{y}}{\partial \mathbf{u}_1}\cdots\frac{\partial \mathbf{y}}{\partial \mathbf{u}_k}\binom{d_i}{k}\langle \mathbf{y},\mathbf{x}\rangle^{d-k}\langle \mathbf{y},\mathbf{u}_1\rangle\cdots\langle \mathbf{y},\mathbf{u}_k\rangle \\
&= \frac{1}{k!}\binom{d_i}{k}\operatorname{Perm}\left[\langle \mathbf{u}_i,\mathbf{u}_j\rangle\right] \\
&\leq \binom{d_i}{k}
\end{aligned}$$

It follows that

$$\frac{1}{k!}\left\|D\mathbf{F}_{\mathbf{x}}(0)^{-1}D^k\mathbf{F}_{\mathbf{x}}(0)\right\| \leq \mu(\mathbf{f},\mathbf{x})\max\frac{1}{\sqrt{d_i}}\binom{d_i}{k}.$$

Estimating $\binom{d_i}{k} \leq d_i^k 2^{-k}$ and using Exercise 8.1,

$$\gamma(\mathbf{F}_{\mathbf{x}},0) \leq \frac{d^{3/2}}{2}\mu(\mathbf{f},\mathbf{x}).$$

□

Whenever the sequence $(\mathbf{X}_k)_{k\in\mathbb{N}}$ defined by $\mathbf{X}_0 = 0$, $\mathbf{X}_{k+1} = = N(\mathbf{F}_{\mathbf{x}},\mathbf{X}_k)$ converges, let $\mathbf{X}^* = \lim \mathbf{X}_k$ and define

$$\zeta_x = \frac{\mathbf{x}+\mathbf{X}^*}{\|\mathbf{x}+\mathbf{X}^*\|} \in S^{n+1}.$$

As in Theorem 5.3, define

$$r_0(\alpha) = \frac{1+\alpha-\sqrt{1-6\alpha+\alpha^2}}{4\alpha}$$

Let α_* the smallest positive root of

$$\alpha_* = \alpha_0(1-\alpha_* r_0(\alpha_*))^2.$$

Numerically, $\alpha_* > 0.116$. (This is better than (Cucker et al., 2008)). Let $B_{\mathbf{x}} = \{\mathbf{y}\in S^n : \rho(\mathbf{x},\mathbf{y}) \leq r_{\mathbf{x}}\}$ with $r_{\mathbf{x}} = r_0(\alpha_*)\mu(\mathbf{f},\mathbf{x})\|\mathbf{f}(\mathbf{x})\|$.

Theorem 9.2. *Let* $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ *and* $\mathbf{x} \in S^n$ *be such that*

$$(\max d_i)^{3/2}\mu(\mathbf{f},\mathbf{x})^2\|\mathbf{f}(\mathbf{x})\| \leq \alpha_*.$$

Then,

(1) $\alpha(\mathbf{F},0) \leq \alpha_*$.
(2) 0 *is an approximate zero of the second kind of* $\mathbf{F}_{\mathbf{x}}$*, and in particular* $\mathbf{f}(\zeta_x) = 0$.
(3) $\zeta_{\mathbf{x}} \in B_{\mathbf{x}}$.
(4) *For any* $\mathbf{z} \in B_{\mathbf{x}}$, $\zeta_{\mathbf{z}} = \zeta_{\mathbf{x}}$.

Proof. (1) By Lemma 9.1,

$$\alpha(\mathbf{F}_{\mathbf{x}},0) \leq (\max d_i)^{3/2}\mu(\mathbf{f},\mathbf{x})\left\|D\mathbf{f}(\mathbf{x})^{-1}_{\mathbf{x}^\perp}\mathbf{f}(\mathbf{x})\right\| \leq$$
$$\leq (\max d_i)^{3/2}\mu(\mathbf{f},\mathbf{x})^2\|\mathbf{f}(\mathbf{x})\| \leq \alpha_*.$$

(2) Since $\alpha_* \leq \alpha$, we can apply Theorem 5.3 to $\mathbf{F}_{\mathbf{x}}$ and 0.
(3) Since 0 is a zero of the second kind for $\mathbf{F}_{\mathbf{x}}$,

$$\mathbf{F}_{\mathbf{x}}(\mathbf{X}^*) = \mathbf{f}(\|\mathbf{x}+\mathbf{X}^*\|\zeta_{\mathbf{x}}) = 0$$

and hence by homogeneity $\mathbf{f}(\zeta_{\mathbf{x}}) = 0$.
(4)

$$\rho(\mathbf{x},\zeta_{\mathbf{x}}) \leq \tan\rho(\mathbf{x},\zeta_{\mathbf{x}}) \leq r_0(\alpha_*)\beta(\mathbf{f},\mathbf{x}) \leq r_0(\alpha_*)\mu(\mathbf{f},\mathbf{x})\|\mathbf{f}(\mathbf{x})\|$$

(5) By Theorem 8.2,

$$\mu(\mathbf{f},\mathbf{z}) \leq \frac{1}{1-(\max d_i)\mu(\mathbf{f},\mathbf{x})\rho(\mathbf{x},\mathbf{z})}\mu(\mathbf{f},\mathbf{x}) \leq \frac{1}{1-\alpha^* r_0(\alpha_*)}\mu(\mathbf{f},\mathbf{x})$$

and hence, as in item 1:

$$\alpha(\mathbf{F}_{\mathbf{z}},0) \leq \frac{1}{(1-\alpha^* r_0(\alpha_*))^2}\alpha_* \leq \alpha_0.$$

□

This theorem appeared in Cucker et al. (2008). For other inclusion/exclusion theorems based in alpha-theory, see Giusti et al. (2007).

10. The exclusion lemma

Lemma 10.1. *Let* $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ *and let* $\mathbf{x},\mathbf{y} \in S^n$ *with* $\rho(\mathbf{x},\mathbf{y}) \leq \sqrt{2}$. *Then,*

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq \max(d_i)\rho(\mathbf{x},\mathbf{y}).$$

In particular, let $\delta = \min(\|\mathbf{f}(\mathbf{x})\|/\sqrt{\max(d_i)},\sqrt{2})$. *If* $\mathbf{f}(\mathbf{x}) \neq 0$, *then there is no zero of* $\mathbf{f}$ *in*

$$B(\mathbf{x},\delta) = \{\mathbf{y} \in S^{n+1} : \rho(\mathbf{x},\mathbf{y}) \leq \delta\}.$$

Proof. First of all,

$$\begin{aligned} |f_i(x) - f_i(y)| &= |\langle f_i(\cdot), K_{d_i}(\cdot, \mathbf{x}) - K_{d_i}(\cdot, \mathbf{y})\rangle| \\ &\leq \|f_i\| \|K_{d_i}(\cdot, \mathbf{x}) - K_{d_i}(\cdot, \mathbf{y})\| \\ &\leq \|f_i\| \sqrt{K_{d_i}(\mathbf{x}, \mathbf{x}) + K_{d_i}(\mathbf{y}, \mathbf{y}) - 2K_{d_i}(\mathbf{x}, \mathbf{y})} \\ &= \|f_i\| \sqrt{2} \sqrt{1 - \cos(\theta)^d} \end{aligned}$$

with $\theta = \rho(x, y)$. Since $\theta \leq \pi < \sqrt{30}$, we have always

$$\cos(\theta) = 1 - \frac{1}{2}\theta^2 + \frac{1}{4!}\theta^4 - \frac{1}{6!}\theta^6 + \cdots > 1 - \frac{1}{2}\theta^2.$$

The reader will check that for $\epsilon < 1$, $(1 - \epsilon)^d > 1 - d\epsilon$. Therefore, using $\theta < 1/\sqrt{2}$,

$$|f_i(\mathbf{x}) - f_i(\mathbf{y})| \leq \|f_i\| \sqrt{d_i}\theta$$

and

$$\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq \sqrt{\max(d_i)}\theta.$$

□

Part 3. The algorithm and its complexity

11. Convexity and geometry Lemmas

Definition 11.1. Let $\mathbf{y}_1, \ldots, \mathbf{y}_s \in S^n$ belong to the same hemisphere, that is $\langle \mathbf{y}_i, \mathbf{z} \rangle > 0$ for a fixed $\mathbf{z}$. The **spherical convex hull** of $\mathbf{y}_1, \ldots, \mathbf{y}_s$ is defined as

$$\mathrm{SCH}(\mathbf{y}_1, \ldots, \mathbf{y}_s) = \left\{ \frac{\lambda_1 \mathbf{y}_1 + \cdots + \lambda_s \mathbf{y}_s}{\|\lambda_1 \mathbf{y}_1 + \cdots + \lambda_s \mathbf{y}_s\|} : \lambda_1, \ldots, \lambda_s \geq 0 \right.$$
$$\left. \text{and } \lambda_1 + \cdots + \lambda_s = 1 \right\}.$$

This is the same as the intersection of the sphere with the cone $\{\lambda_1 \mathbf{y}_1 + \cdots + \lambda_s \mathbf{y}_s : \lambda_1, \ldots, \lambda_s \geq 0\}$. We will need the following convexity Lemma from Cucker et al. (2008):

Lemma 11.2. *Let $\mathbf{y}_1, \ldots, \mathbf{y}_s \in S^n$ belong to the same hemisphere. Let $r_1, \ldots, r_s > 0$ and let $B(\mathbf{y}_i, r_i) = \{\mathbf{x} \in S^n : \rho(x, \mathbf{y}_i) < r_i\}$. If $\cap B(\mathbf{y}_i, r_i) \neq \emptyset$, then $\mathrm{SCH}(\mathbf{y}_1, \ldots, \mathbf{y}_s) \subset \cup B(\mathbf{y}_i, r_i)$.*

Exercise 11.1. Prove Lemma 11.2 above.

For the root counting algorithm, we will need to define a **mesh** on the sphere.

Lemma 11.3. *For every $\eta = 2^{-t}$, we can construct a set $C(\eta) \subseteq S^n$ satisfying:*

1. *For all $\mathbf{z} \in S^n, \exists \mathbf{x} \in C(\eta)$ such that $\rho(\mathbf{z}, \mathbf{x}) \leq \eta\sqrt{n}/2$.*
2. *For all $\mathbf{x} \in S^n$, let $Y = \{\mathbf{y} \in C(\eta) : \rho(\mathbf{x}, \mathbf{y}) \leq \sqrt{n}\eta\}$. Then $\mathbf{x} \in \mathrm{SCH}(Y)$.*
3. *$\#C(\eta) \leq 2n(1 + 2^{t+1})^n$.*

Proof. Just set

$$C(\eta) = \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{R}^{n+1}, x_i\eta^{-1} \in \mathbb{Z}, \|\mathbf{x}\|_\infty = 1 \right\}.$$

This corresponds to dividing $Q = \{\mathbf{x} : \|\mathbf{x}\|_\infty = 1\}$ into n-cubes of side $\tilde{\eta}$. The maximal distance in Q between a point $\mathbf{Z} \in Q$ and a point $\mathbf{X}$ in the mesh is half of the diagonal, or $\eta\sqrt{n}$. Then

$$\rho(\mathbf{Z}/\|\mathbf{Z}\|, \mathbf{X}/\|\mathbf{X}\|) < \eta\sqrt{n}.$$

Now, let Y' be the set of points $\mathbf{y} \in C(\eta)$ such that the distance along Q between $\mathbf{x}/\|\mathbf{x}\|_\infty$ and $\mathbf{y}/\|\mathbf{y}\|_\infty$ is at most η. Then clearly $\mathbf{x} \in \mathrm{SCH}(Y')$. Moreover, $Y' \subset Y$.

The last item is trivial. □

12. The counting algorithm

Given $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ and $\eta = 2^{-t}$, we construct a graph $\mathcal{G}_\eta = (\mathcal{V}_\eta, \mathcal{E}_\eta)$ as follows. Let

$$A(\mathbf{f}) = \{\mathbf{x} \in S^n : \max d_i^{3/2}\mu(\mathbf{f}, \mathbf{x})^2\|\mathbf{f}(\mathbf{x})\| < \alpha_*\}$$

be the set of points satisfying the hypotheses of Theorem 9.2. The set of vertices of $\mathcal{G}_\eta$ is $\mathcal{V}_\eta = C(\eta) \cap A(\mathbf{f})$.

Recall that Let $B_{\mathbf{x}} = \{\mathbf{y} \in S^n : \rho(\mathbf{x}, \mathbf{y}) \leq r_{\mathbf{x}}\}$ with $r_{\mathbf{x}} = r_0(\alpha_*)\mu(\mathbf{f}, \mathbf{x})\|\mathbf{f}(\mathbf{x})\|$. The set of edges of $\mathcal{G}_\eta$ is $\mathcal{E}_\eta = \{(\mathbf{x}, \mathbf{y}) \in \mathcal{V}_\eta \times \mathcal{V}_\eta : B_{\mathbf{x}} \cap B_{\mathbf{y}} \neq \emptyset\}$. This graph is clearly constructible. Theorem 9.2 implies that for any edge $(\mathbf{x}, \mathbf{y}) \in \mathcal{E}_\eta$, $\zeta_{\mathbf{x}} = \zeta_{\mathbf{y}}$. More generally,

Lemma 12.1. *The vertices of any connected component of $\mathcal{G}(\eta)$ are approximate zeros associated to the same zero of $\mathbf{f}$. Moreover, if $\mathbf{x}, \mathbf{y}$ belong to distinct connected components of $\mathcal{G}(\eta)$, then $\zeta_{\mathbf{x}} \neq \zeta_{\mathbf{y}}$.*

The algorithm is as follows:

Algorithm `RootCount`
Input: $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$.
Output: $\#\zeta \in S^n : \mathbf{f}(\zeta) = 0$.

$\eta \leftarrow 2^{-\lceil \log_2(1/\sqrt{2n}) \rceil}$.

Repeat

$\eta \leftarrow \eta/2$.

Let $\mathcal{U}_1, \dots, \mathcal{U}_r$ be the connected components of $\mathcal{G}_\eta$.

Until $\forall 1 \leq i < j \leq r, \forall \mathbf{x}$ vertex of $\mathcal{U}_i, \forall \mathbf{y}$ vertex of $\mathcal{U}_j$,

$$\rho(\mathbf{x}, \mathbf{y}) > 2\eta\sqrt{n}. \tag{16}$$

and $\forall \mathbf{x} \in C(\eta) \setminus A(\mathbf{f})$,

$$\|\mathbf{f}(\mathbf{x})\| > \eta\sqrt{n \max d_i}/2. \tag{17}$$

Return r.

Theorem 12.2. *If the algorithm* RootCount *stops, then* r *is the correct number of roots of* $\mathbf{f}$ *in* S^n.

Proof of Th.12.2. Suppose the algorithm stopped at a certain value of η. As each connected component $\mathcal{U}_i$ determines a distinct and unique zero of $\mathbf{f}$, it remains to prove that there are no zeros of $\mathbf{f}$ outside $\cup_{\mathbf{x} \in \mathcal{V}_\eta} B_{\mathbf{x}}$.

Therefore, assume by contradiction that there is $\zeta \in S^n$ with $\mathbf{f}(\zeta) = 0$ and $\zeta \notin B_{\mathbf{x}}$ for any $\mathbf{x} \in V_\eta$.

Let Y be the set of $\mathbf{y} \in C(\eta)$ with $\rho(\zeta, \mathbf{y}) \leq \eta\sqrt{n}$.

If there is $\mathbf{y} \in Y$ with $\mathbf{y} \notin A(\mathbf{f})$ let $\delta = \|\mathbf{f}(\mathbf{y})\|/\sqrt{\max d_i}$. Equation (17) guarantees that $\eta\sqrt{n}/2 < \delta$. By construction, $\eta\sqrt{n}/2 < \sqrt{2}$. Therefore, the exclusion lemma 10.1 guarantees that $\mathbf{f}(\zeta) \neq 0$, contradiction.

Therefore, we assume that $Y \subset A(\mathbf{f})$. Equation (16) guarantees that $Y \subset \mathcal{U}_k$ for a same connected component of $\mathcal{G}_\eta$. Therefore, $\cap_{\mathbf{y} \in Y} B_{\mathbf{y}} \ni \zeta$ is not empty.

By Lemma 11.3(2), $\mathbf{x} \in \mathrm{SCH}(Y)$. Lemma 11.2 says that

$$\mathrm{SCH}(Y) \subseteq \cup_{\mathbf{y} \in Y} B_{\mathbf{y}}$$

Thus, $\mathbf{x} \in B_{\mathbf{y}}$ for some $\mathbf{y}$, contradiction again.

□

A consequence of Th.12.2 is that **if the algorithm stops**, one can obtain an approximate zeros of the second kind for each root of f by recovering one vertex for each connected component.

13. Complexity

We did not prove that algorithm RootCount stops. It actually stops **almost surely**, that is for input f outside a certain measure zero set.

Define

$$\kappa(\mathbf{f}, \mathbf{x}) = \frac{1}{\sqrt{\mu(\mathbf{f}, \mathbf{x})^{-2} + \|\mathbf{f}(\mathbf{x})\|^2}}$$

and notice that

$$\kappa(\mathbf{f},\mathbf{x}) \leq \mu(\mathbf{f},\mathbf{x}) \text{ and } \kappa(\mathbf{f},\mathbf{x}) \leq \|\mathbf{f}(\mathbf{x})\|^{-1}.$$

Reciprocally,

$$\min(\mu(\mathbf{f},\mathbf{x}), \|\mathbf{f}(\mathbf{x})\|^{-1}) \leq \sqrt{2}\kappa(\mathbf{f},\mathbf{x}).$$

If $\mathbf{f}(\mathbf{x}) = 0$, then $\kappa(\mathbf{f},\mathbf{x}) = \mu(\mathbf{f},\mathbf{x})$.

Definition 13.1. The **condition number** for for Problem 1.2 (counting real zeros on the sphere) is

$$\kappa(\mathbf{f}) = \max_{\mathbf{x}\in S^n} \kappa(\mathbf{f},\mathbf{x}).$$

Assume that $\mathbf{f}$ has no degenerate root. Then the denominator is bounded away from zero, and $\kappa(\mathbf{f})$ is finite. We will prove later that the algorithm stops for $\kappa(\mathbf{f})$ finite. But before, we state and prove the **condition number theorem** to obtain some geometric intuition on $\kappa(\mathbf{f})$.

Theorem 13.2. *(Cucker, Krick, Malajovich, and Wschebor, 2009) Let* $\Sigma^{\mathbb{R}} = \{\mathbf{g} \in \mathcal{H}_{\mathbf{d}}^{\mathbb{R}} : \exists \zeta \in S^n : \mathbf{g}(\zeta) = 0 \text{ and } \operatorname{rk}(D\mathbf{g}(\zeta)) < n\}$. *Let* $\mathbf{f} \in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$, $\mathbf{f} \notin \Sigma^{\mathbb{R}}$. *Then,*

$$\kappa(\mathbf{f}) = \frac{1}{\min_{\mathbf{g}\in\Sigma^{\mathbb{R}}} \|\mathbf{f}-\mathbf{g}\|}.$$

In particular, $\kappa(\mathbf{f}) \geq 1$.

Proof. It suffices to prove that

$$\kappa(\mathbf{f},\mathbf{x}) = \frac{1}{\min\limits_{\substack{\mathbf{g}\in\mathcal{H}_{\mathbf{d}}^{\mathbb{R}} \\ \mathbf{g}(\mathbf{x})=0 \\ \operatorname{rk}(D\mathbf{g}(\mathbf{x}))<n}} \|\mathbf{f}-\mathbf{g}\|}.$$

We proceed as in the proof of Prop.8.1. We decompose

$$\mathcal{H}_{\mathbf{d}}^{\mathbb{R}} = H_0 \oplus H_1 \oplus H_2 \oplus \cdots$$

where H_0 and H_1 correspond to the constant and linear terms of $\mathbf{y} \mapsto \mathbf{f}(\mathbf{x}+\mathbf{y})$. Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be an orthonormal basis for $\mathbf{x}^\perp$.

An orthonormal basis for $H_0 \oplus H_1$ is

$$\left(K_{d_i}(\cdot,\mathbf{x}), \frac{1}{\sqrt{d}} \frac{\partial K_{d_i}(\cdot,\mathbf{x})}{\partial \mathbf{u}_j} \right).$$

The projection of $\mathbf{f}$ in $H_0 \oplus H_1$ is

$$\left[\langle \mathbf{f}(\cdot), K_{d_i}(\cdot, \mathbf{x})\rangle\right] \oplus \begin{bmatrix} & \vdots & \\ \cdots & \left\langle \mathbf{f}_i, \frac{1}{\sqrt{d}} \frac{\partial K_{d_i}(\cdot,\mathbf{x})}{\partial \mathbf{u}_j} \right\rangle & \cdots \\ & \vdots & \end{bmatrix} =$$

$$= \mathbf{f}(\mathbf{x}) \oplus \begin{bmatrix} d_1^{-1/2} & & \\ & d_2^{-1/2} & \\ & & d_n^{-1/2} \end{bmatrix} D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp}.$$

This is an orthogonal projection onto $\mathbb{R}^n \times \mathbb{R}^{n\times n}$.

Now,

$$\kappa(\mathbf{f}, \mathbf{x})^{-2} = \|\mathbf{f}(\mathbf{x})\|^2 + \sigma_n \left(\begin{bmatrix} d_1^{-1/2} & & \\ & d_2^{-1/2} & \\ & & d_n^{-1/2} \end{bmatrix} D\mathbf{f}(\mathbf{x})_{|\mathbf{x}^\perp} \right).$$

Again, we apply Th.6.3. □

Lemma 13.3. *Let ζ_1, ζ_2 be distinct roots of $\mathbf{f}$ in S^n. Then,*

$$\rho(\zeta_1, \zeta_2) \geq \frac{1}{\max d_i^{3/2} \kappa(\mathbf{f})}$$

Proof.

$$\begin{aligned} \|\zeta_1 - \zeta_2\| &\geq \frac{1}{2\gamma(\mathbf{f}, \zeta_1)} && \text{by Ex.4.3} \\ &\geq \frac{1}{\max d_i^{3/2} \mu(\mathbf{f}, \zeta_1)} && \text{by Lem.9.1} \\ &\geq \frac{1}{\max d_i^{3/2} \kappa(\mathbf{f})} && \text{because } \mathbf{f}(\zeta_1) = 0. \end{aligned}$$

The Lemma follows. □

Lemma 13.4. *Assume that*

$$\eta < \frac{1}{2 \max d_i^{3/2} \sqrt{n} \kappa(\mathbf{f})} (1 - 2\alpha_* r_0(\alpha_*)).$$

Then (16) *holds.*

Proof. Recall that $\mathbf{x}$ and $\mathbf{y}$ belong to $A_{\mathbf{f}}$, so that

$$\max d_i^{3/2} \mu(\mathbf{f}, \mathbf{x})^2 \|\mathbf{f}(\mathbf{x})\| < \alpha_*$$

and the same for $\mathbf{y}$. In particular, the radius $r_{\mathbf{x}}$ of $B_{\mathbf{x}}$ satisfies

$$r_0(\alpha_*)\mu(\mathbf{f},\mathbf{x})\|\mathbf{f}(\mathbf{x})\| < \frac{\alpha_* r_0(\alpha_*)}{\max d_i^{3/2}\mu(\mathbf{f},\mathbf{x})} \leq \frac{\alpha_* r_0(\alpha_*)}{\max d_i^{3/2}\kappa(\mathbf{f},\mathbf{x})}.$$

By Lemma 13.3 and the triangle inequality,

$$\begin{aligned}\rho(\mathbf{x},\mathbf{y}) &\geq \rho(\zeta_{\mathbf{x}},\zeta_{\mathbf{y}}) - r_0(\alpha_*)\mu(\mathbf{f},\mathbf{x})\|\mathbf{f}(\mathbf{x})\| - r_0(\alpha_*)\mu(\mathbf{f},\mathbf{y})\|\mathbf{f}(\mathbf{y})\| \\ &\geq \frac{1}{\max d_i^{3/2}\kappa(\mathbf{f})}(1-2\alpha_* r_0(\alpha_*)).\end{aligned}$$

□

Lemma 13.5. *Let $\mathbf{x} \notin A_f$. Then,*

$$\|\mathbf{f}(\mathbf{x})\| \geq \frac{\alpha_*}{\kappa(\mathbf{f},\mathbf{x})^2 \max d_i^{3/2}}.$$

Proof. Let $\mathbf{x} \notin A_{\mathbf{f}}$, so that

$$\frac{\max d_i^{3/2}}{2}\mu(\mathbf{f},\mathbf{x})^2\|\mathbf{f}(\mathbf{x})\| \geq \alpha_*.$$

Recall that

$$\min(\mu(\mathbf{f},\mathbf{x}), \|\mathbf{f}(\mathbf{x})\|^{-1}) \leq \sqrt{2}\kappa(\mathbf{f},\mathbf{x})$$

There are two possibilities. If $\mu(\mathbf{f},\mathbf{x}) \leq \sqrt{2}\kappa(\mathbf{f},\mathbf{x})$, then

$$\|\mathbf{f}(\mathbf{x})\| \geq \frac{\alpha_*}{\max d_i^{3/2}\kappa(\mathbf{f},\mathbf{x})^2}.$$

Otherwise,

$$\|\mathbf{f}(\mathbf{x})\| \geq \frac{1}{\sqrt{2}\kappa(\mathbf{f},\mathbf{x})} \geq \frac{\alpha_*}{\max d_i^{3/2}\kappa(\mathbf{f},\mathbf{x})^2}.$$

□

Now we can state the 'cloud complexity' theorem.

Theorem 13.6. *The algorithm* `RootCount` *will stop for*

$$\eta < \frac{1}{\max d_i^{3/2}\kappa(\mathbf{f})^2}\min\left(\alpha_*\ ,\ \frac{\kappa(\mathbf{f})}{2\sqrt{n}}(1-2\alpha_* r_0(\alpha_*))\right)$$

that is, after $O(\log\kappa(\mathbf{f}) + \log\max d_i)$ iterations. The total number of evaluations of $\mathbf{f}$ and $D\mathbf{f}$ is

$$2n(1+4\max d_i^{3/2}\sqrt{n}\kappa(\mathbf{f})^2)^n.$$

That means that $2n(1+4\max d_i^{3/2}\sqrt{n}\kappa(\mathbf{f})^2)^n$ processors in parallel can compute the root count in time $O(\log\kappa(\mathbf{f})+\log\max d_i)$ times a polynomial in n for the linear algebra.

For people concerned with the overall computing cost, a price tag exponential in n is known as the **curse of dimensionality**. It usually plagues divide and conquer and Monte-Carlo algorithms.

But the situation $n=2$ is already interesting. How efficiently can we count zeros of a system of polynomials on the 2-sphere? As the parallel and sequential running time depends upon $\kappa(f)$, it is useful to known more about the condition number.

14. Probabilistic and smoothed analysis

One possibility is to pick the input system $\mathbf{f}$ at random, and treat $\kappa(\mathbf{f})$ as a random variable. For instance, let $\mathbf{f}\in\mathcal{H}_{\mathbf{d}}^{\mathbb{R}}$ be random with **Gaussian** probability distribution

$$\frac{1}{(2\pi)^{\dim\mathcal{H}_{\mathbf{d}}^{\mathbb{R}}/2}}e^{-\|f\|^2/2}\,\mathrm{d}\mathcal{H}_{\mathbf{d}}^{\mathbb{R}}.$$

The tail for the random variable $\kappa(\mathbf{f})$ and the expected value of $\log\kappa(\mathbf{f})$ can be bounded by

Theorem 14.1. *(Cucker, Krick, Malajovich, and Wschebor, 2012) Let $\mathbf{f}$ be as above. Assume that $n\geq 3$. Then,*

(i) *For $a>4\sqrt{2}\,(\max d_i)^2 n^{7/2}N^{1/2}$ we have*

$$\mathrm{Prob}\big(\kappa(\mathbf{f})>a\big)\leq K_n\frac{\sqrt{2n}(1+\ln(a/\sqrt{2n}))^{1/2}}{a},$$

where $N=\dim\mathcal{H}_{\mathbf{d}}^{\mathbb{R}}$, $K_n:=8(\max d_i)^2\mathcal{D}^{1/2}\,N^{1/2}n^{5/2}+1$ and $\mathcal{D}=\prod d_i$.

(ii)

$$\mathbb{E}(\ln\kappa(\mathbf{f}))\leq\ln K_n+(\ln K_n)^{1/2}+(\ln K_n)^{-1/2}+\frac{1}{2}\ln(2n).$$

Notice as a consequence that the expected running time of `RootCount` is $\mathbb{E}(\ln\kappa(\mathbf{f}))\in\mathcal{O}(n\ln\max d_i)$. This is cloud computing time, of course.

Average time analysis depends upon an arbitrary distribution. Spielman and Teng (2004) suggested looking instead at a small random perturbation for each given input. This is known as **smoothed analysis**.

For a given $\mathbf{f}\in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$, we will consider the uniform distribution in the ball $B(\mathbf{f},\arcsin\sigma)\subset S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})$ where σ is an arbitrary radius, and Riemannian metric on the sphere is assumed. The strange looking arcsine comes from the fact that $B(\mathbf{f},\arcsin\sigma)$ is the projection on the

sphere of the ball $B(\mathbf{f}, \sigma) \subset \mathcal{H}_{\mathbf{d}}^{\mathbb{R}}$. The reason for looking at the uniform distribution for perturbations instead of Gaussian is the following result:

Theorem 14.2. *(Bürgisser, Cucker, and Lotz, 2008) Let $\Sigma \subset \mathbb{R}^N$ be contained in a projective hypersurface H of degree at most D and let $\kappa : \mathbb{S}^{N-1} \to [1, \infty]$ be given by*

$$\kappa(\mathbf{f}) = \frac{\|\mathbf{f}\|}{\min_{\mathbf{g}\in\Sigma} \|\mathbf{f} - \mathbf{g}\|}.$$

Then, for all $\sigma \in (0, 1]$,

$$\sup_{\mathbf{f}\in S^{N-1}} \mathrm{E}_{\mathbf{h}\in B(\mathbf{f},\arcsin\sigma)\subseteq S^{N-1}}(\ln \kappa(\mathbf{h})) \leq 2\ln(N-1) + 2\ln D - \ln\sigma + 5.5.$$

In the context of the root counting problem, the degree D of $\Sigma = \Sigma^{\mathbb{R}}$ is bounded by $n^2(\prod d_i)(\max d_i)$. Therefore,

Corollary 14.3. *(Cucker, Krick, Malajovich, and Wschebor, 2009)*

$$\sup_{\mathbf{f}\in S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})} \mathrm{E}_{\mathbf{h}\in B(\mathbf{f},\arcsin\sigma)\subseteq S(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})}(\ln \kappa(h)) \leq 2\ln(\dim(\mathcal{H}_{\mathbf{d}}^{\mathbb{R}})) + 4\ln(n) + 2\ln(\prod d_i) + \ln 1/\sigma + 6.$$

15. Conclusions

We sketched the average time analysis and a smoothed analysis of an algorithm for real root counting and, incidentally, root finding. The same algorithm can also decide if a given polynomial system admits a root.

Loosely speaking, deciding (resp. counting) roots of polynomial systems are NP-complete (resp. #P complete) problems. The formal NP-complete and #P-complete problems refer to **sparse** polynomial systems.

Our algorithm requires actually polynomial evaluations, so it can take advantage of the sparse structure. Moreover, the degree of the sparse discriminant is no more than the degree of the usual discriminant. In that sense Corollary 14.3 is still valid. The running time of the algorithm is polynomial in n and in the dimension of the input space. Again, this is a massively parallel algorithm so the number of processors is exponential in n.

References

Blum, Lenore, Felipe Cucker, Michael Shub, and Steve Smale. 1998. *Complexity and real computation*, Springer-Verlag, New York. With a foreword by Richard M. Karp. MR1479636 (99a:68070)

Bürgisser, Peter and Felipe Cucker. 2006. *Counting complexity classes for numeric computations. II. Algebraic and semialgebraic sets*, J. Complexity **22**, no. 2, 147–191, DOI 10.1016/j.jco.2005.11.001. MR2200367 (2007b:68059)

_____. 2011. *On a problem posed by Steve Smale*, Ann. of Math. (2) **174**, no. 3, 1785–1836, DOI 10.4007/annals.2011.174.3.8. MR2846491

Bürgisser, Peter, Felipe Cucker, and Martin Lotz. 2008. *The probability that a slightly perturbed numerical analysis problem is difficult*, Math. Comp. **77**, no. 263, 1559–1583, DOI 10.1090/S0025-5718-08-02060-7. MR2398780 (2009a:65132)

Cucker, Felipe, Teresa Krick, Gregorio Malajovich, and Mario Wschebor. 2008. *A numerical algorithm for zero counting I: Complexity and accuracy*, Journal of Complexity **24**, no. 5-6, 582-605, DOI 10.1016/j.jco.2008.03.001.

_____. 2009. *A numerical algorithm for zero counting II: Distance to Ill-posedness and Smoothed Analysis*, Journal of Fixed Point Theory and Applications **6**, no. 2, 285-294, DOI 10.1007/s11784-009-0127-4.

_____. 2012. *A numerical algorithm for zero counting. III: Randomization and condition*, Adv. in Appl. Math. **48**, no. 1, 215–248, DOI 10.1016/j.aam.2011.07.001. MR2845516

Dedieu, Jean-Pierre. 1997a. *Estimations for the separation number of a polynomial system*, J. Symbolic Comput. **24**, no. 6, 683–693, DOI 10.1006/jsco.1997.0161. MR1487794 (99b:65065)

_____. 1997b. *Estimations for the separation number of a polynomial system*, J. Symbolic Comput. **24**, no. 6, 683–693, DOI 10.1006/jsco.1997.0161. MR1487794 (99b:65065)

Dedieu, Jean-Pierre, Gregorio Malajovich, and Mike Shub. 2012. *Adaptive Step Size Selection for Homotopy Methods to Solve Polynomial Equations*, IMA Journal of Numerical Analysis. `http://dx.doi.org/doi:10.1093/imanum/drs007`.

Demmel, James W. 1988. *The probability that a numerical analysis problem is difficult*, Math. Comp. **50**, no. 182, 449–480, DOI 10.2307/2008617. MR929546 (89g:65062)

Edelman, Alan. 1992. *On the distribution of a scaled condition number*, Math. Comp. **58**, no. 197, 185–190, DOI 10.2307/2153027. MR1106966 (92g:15034)

Eckart, Carl and Gale Young. 1936. *The approximation of a matrix by another of lower rank*, Psychometrika **1**, no. 3, 211–218, DOI 10.1007/BF02288367.

_____. 1939. *A principal axis transformation for non-hermitian matrices*, Bull. Amer. Math. Soc. **45**, no. 2, 118–121, DOI 10.1090/S0002-9904-1939-06910-3.

Giusti, M., G. Lecerf, B. Salvy, and J.-C. Yakoubsohn. 2007. *On location and approximation of clusters of zeros: case of embedding dimension one*, Found. Comput. Math. **7**, no. 1, 1–49, DOI 10.1007/s10208-004-0159-5. MR2283341 (2008e:65159)

Higham, Nicholas J. 2002. *Accuracy and stability of numerical algorithms*, 2nd ed., Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR1927606 (2003g:65064)

The Institute of Electrical and Electronics Engineers Inc. 2008. *IEEE Standard for Floating Point Arithmetic IEEE Std 754-2008*, 3 Park Avenue, New York, NY 10016-5997, USA, `http://ieeexplore.ieee.org/xpl/standards.jsp`.

Kantorovich, L. V. 1949. *On the Newton method*, in: L.V. Kantorovich, *Selected works. Part II, Applied functional analysis. Approximation methods and computers;*, Classics of Soviet Mathematics, vol. 3, Gordon and Breach Publishers, Amsterdam, 1996. Translated from the Russian by A. B. Sossinskii; Edited by S. S. Kutateladze and J. V. Romanovsky. Article originally published in *Trudy MIAN SSSR* **28** 104-144(1949).

Malajovich, Gregorio. 1993. *On the complexity of path-following Newton algorithms for solving systems of polynomial equations with integer coefficients*, PhD Thesis, Department of Mathematics, University of California at Berkeley, `http://www.labma.ufrj.br/~gregorio/papers/thesis.pdf`.

_____. 1994. *On generalized Newton algorithms: quadratic convergence, path-following and error analysis*, Theoret. Comput. Sci. **133**, no. 1, 65–84, DOI 10.1016/0304-3975(94)00065-4. Selected papers of the Workshop on Continuous Algorithms and Complexity (Barcelona, 1993). MR1294426 (95g:65073)

_____. 2011. *Nonlinear equations*, Publicações Matemáticas do IMPA. [IMPA Mathematical Publications], Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro. With an appendix by Carlos Beltrán, Jean-Pierre Dedieu, Luis Miguel Pardo and Mike Shub; 28º Colóquio Brasileiro de Matemática. [28th Brazilian Mathematics Colloquium]. MR2798351 (2012j:65148)

Meer, Klaus. 2000. *Counting problems over the reals*, Theoret. Comput. Sci. **242**, no. 1-2, 41–58, DOI 10.1016/S0304-3975(98)00190-X. MR1769145 (2002g:68041)

Nachbin, Leopoldo. 1964. *Lectures on the Theory of Distributions*, Textos de Matemática, Instituto de Física e Matemática, Universidade do Recife.

_____. 1969. *Topology on spaces of holomorphic mappings*, Ergebnisse der Mathematik und ihrer Grenzgebiete, Band 47, Springer-Verlag New York Inc., New York. MR0254579 (40 #7787)

Shub, Michael and Steve Smale. 1993. *Complexity of Bézout's theorem. I. Geometric aspects*, J. Amer. Math. Soc. **6**, no. 2, 459–501, DOI 10.2307/2152805. MR1175980 (93k:65045)

Smale, Steve. 1985. *On the efficiency of algorithms of analysis*, Bull. Amer. Math. Soc. (N.S.) **13**, no. 2, 87–121, DOI 10.1090/S0273-0979-1985-15391-1. MR799791 (86m:65061)

_____. 1986. *Newton's method estimates from data at one point*, computational mathematics (Laramie, Wyo., 1985), Springer, New York, pp. 185–196. MR870648 (88e:65076)

Spielman, Daniel A. and Shang-Hua Teng. 2004. *Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time*, J. ACM **51**, no. 3, 385–463 (electronic), DOI 10.1145/990308.990310. MR2145860 (2006f:90029)

Turing, A. M. 1948. *Rounding-off errors in matrix processes*, Quart. J. Mech. Appl. Math. **1**, 287–308. MR0028100 (10,405c)

Wang Xinghua. 1993. *Some result relevant to Smale's reports*, in: M.Hirsch, J. Marsden and S. Shub(eds): *From Topolgy to Computation: Proceedings of Smalefest*, Springer, new-york, pp. 456-465.

Wilkinson, J. H. 1994. *Rounding errors in algebraic processes*, Dover Publications Inc., New York. Reprint of the 1963 original [Prentice-Hall, Englewood Cliffs, NJ; MR0161456 (28 #4661)]. MR1280465

Departamento de Matemática Aplicada, Instituto de Matemática, Universidade Federal do Rio de Janeiro. Caixa Postal 68530, Rio de Janeiro RJ 21941-909, Brasil.

E-mail address: gregorio.malajovich@gmail.com

URL: www.labma.ufrj.br/~gregorio